UNIVERSIDAD DE CUENCA

FACULTAD DE INGENIERÍA

MAESTRÍA EN GESTIÓN ESTRATÉGICA DE TECNOLOGÍAS DE LA INFORMACIÓN

“Proceso para el descubrimiento de conocimiento en las bases de datos de la Universidad de Cuenca mediante técnicas de Data Mining”

TESIS PREVIA A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN GESTIÓN ESTRATÉGICA DE TECNOLOGÍAS DE LA INFORMACIÓN

AUTOR:

Ing. Gustavo Eduardo Cordero Páez

CI. 01014130091

DIRECTOR:

Ing. Víctor Hugo Saquicela Galarza, PhD

CI. 0103599577

CUENCA - ECUADOR

2017

RESUMEN

Considerando el gran volumen de datos que la Universidad de Cuenca ha recolectado por varias décadas en sus sistemas de información, el presente trabajo pretende facilitar el aprovechamiento de este valioso recurso a través de la definición de un proceso específico para el desarrollo de proyectos de Minería de Datos dentro de esta institución educativa; el cual basándose en el estándar formal de la industria (CRISP-DM) propone un enfoque de metodología ágil, y propone la utilización del sistema de Data Warehouse de la Universidad como principal fuente de acceso a los datos. Dicho proceso establecido se ha validado con su aplicación a tres problemas específicos: a) identificación de los parámetros que se encuentran más relacionados al éxito académico en alumnos nuevos, b) construcción de una herramienta o modelo que permita predecir la deserción estudiantil en alumnos nuevos y c) análisis de grupos sobre la relación existente entre la calificación estudiantil y la evaluación al docente. Como resultado se evidenció que la aplicación de estas técnicas son un aporte contundente a la toma de decisiones basada en los datos (Data-Driven Decision Making), puesto que los resultados obtenidos en cada caso han generado un conocimiento novedoso sobre las problemáticas planteadas, confirmándose además que el proceso definido es una importante ayuda para el desarrollo organizado del proyecto, generando la suficiente documentación para su réplica y análisis posterior.

Palabras clave: Proceso de Minería de Datos, Minería de Datos Educacionales, Analítica de Aprendizaje, CRISP-DM, Descubrimiento de Conocimiento en Bases de Datos.

ABSTRACT

Considering the large volume of data that the University has collected for several decades in its information systems, the present work proposes the use of this valuable resource through the definition of a specific process for the development of Data Mining projects within of the University of Cuenca; which, based on the formal industry standard (CRISP-DM), proposes an agile methodology approach, and seeks to use the University's Data Warehouse system as the main source for data access. This established process has been validated with its application to three specific problems: a) identification of the parameters that are most related to academic success in new students, b) construction of a model to predict student dropout in new students and c) group analysis on the relationship between student qualification and teacher evaluation. As a result, it was confirmed that the application of these techniques are a decisive contribution to Data-Driven Decision Making. On the other hand, it was verified the results obtained in each problem have generated a new knowledge in each case, besides that the process has been defined as an important aid for the organized

development of the project, generating sufficient documentation for its replication and later analysis.

KEYWORDS: Data Mining Process, Educational Data Mining, Learning Analytics, CRISP-DM, Knowledge Discovery in Databases

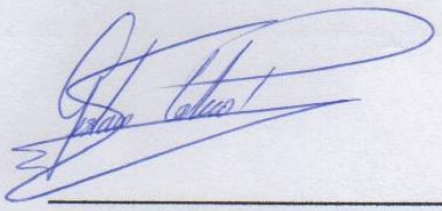
Índice

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

GUSTAVO EDUARDO CORDERO PÁEZ en calidad de autor y titular de los derechos morales y patrimoniales del trabajo de titulación **"Proceso para el descubrimiento de conocimiento en las bases de datos de la Universidad de Cuenca mediante técnicas de Data Mining"**, de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 14 de noviembre de 2017



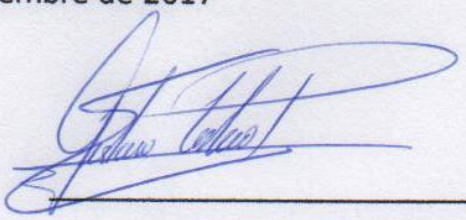
Ing. Gustavo Eduardo Cordero Páez

C.I: 0104130091

Cláusula de Propiedad Intelectual

GUSTAVO EDUARDO CORDERO PÁEZ, autor del trabajo de titulación **"Proceso para el descubrimiento de conocimiento en las bases de datos de la Universidad de Cuenca mediante técnicas de Data Mining"**, certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor/a.

Cuenca, 14 de noviembre de 2017



Ing. Gustavo Eduardo Cordero Páez

C.I: 0104130091

DEDICATORIA

Dedico este esfuerzo de superación profesional y personal a mis hijos y mi esposa, compañeros en esta travesía de sacrificio, aprendizaje y crecimiento, a ellos que son la mayor recompensa, gracias por su apoyo, comprensión y paciencia.

AGRADECIMIENTOS

Este proyecto no podría haber sido desarrollado sin la valiosa colaboración del equipo de la DTIC, agradezco especialmente al Ing. Víctor Saquicela quien como Director del Departamento y del proyecto ha dado las facilidades para que este pueda ser desarrollado de la mejora manera contando siempre con su apoyo y acertada guía. Del mismo modo mi agradecimiento a la Ing. Carmen Rojas y Gustavo Mora por su valiosa colaboración a la hora de indagar y obtener los datos requeridos para el presente análisis.

Un agradecimiento especial a la empresa RapidMiner, quien facilitó una licencia educativa de su herramienta RapidMiner Studio 7.6, utilizada para el desarrollo del tercer problema de minería de datos desarrollado en el capítulo 5.

1 INTRODUCCIÓN

Con el avance en la informatización tanto de las organizaciones como de los individuos, se ha generado una producción sin precedentes de datos, actualmente los datos no sólo reposan en los sistemas transaccionales, sino han aparecido nuevos tipos de sistemas más avanzados que no sólo capturan datos a través de formularios de registro, sino que los obtienen de la monitorización de los usuarios y la forma en que ellos interactúan con los sistemas; almacenando datos de sus hábitos, intereses, experiencias, etc. En la actualidad se dispone de un sin número de fuentes de datos correspondiente a un mismo actor (persona, organización, o cosa), es decir datos que describen el mismo objeto desde perspectivas diferentes, pero, ¿qué sucede si ciertos datos están relacionados entre ellos?, ¿se puede obtener conocimiento de estos datos aparentemente independientes? Este tipo de preguntas fueron ya planteadas hace varias décadas por estadistas que buscaban encontrar correlaciones en los datos disponibles, y que posteriormente a inicio de los años 80, con el avance en informática y el aparecimiento de los sistemas de gestión de bases de datos DBMS (Database Management Systems), se acuñan los términos KDD las siglas en inglés para Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases), y DM las siglas en inglés de Minería de Datos (Data Mining). Ambos términos hacen referencia al “proceso de extracción no trivial de información implícita, previamente desconocida y potencialmente útil”, definición de Minería de Datos dada por William J. Frawley. La Minería de Datos puede ser considerada como parte de la evolución natural desde los sistemas DBMS, la industria de los sistemas de bases de datos han evolucionado desde funcionalidades básicas como el almacenamiento, acceso y procesamiento de los datos, hasta funcionalidades avanzadas como el análisis de los datos (Data Warehousing y Minería de Datos) (Jiawei Han, 2006).

El descubrimiento de conocimiento, al que hacen referencia los términos KDD y Minería de Datos, se refiere a la identificación de patrones en los datos (Yoandry Pacheco, 2015), y se lo realiza a través de un proceso que inicia con la comprensión del negocio a estudiar y los datos disponibles (Chapman, Pete y otros, 2000), luego contempla la obtención y depuración de estos datos para que sean analizados a través de la aplicación de diversas técnicas o algoritmos que involucran varias áreas de conocimiento como la estadística, machine learning y bases de datos; como resultado de este proceso de análisis, se obtienen modelos matemáticos, patrones descubiertos, asociación entre variables, etc., es decir información que al ser analizada e interpretada trae a la luz un nuevo conocimiento, el mismo que permitiría planear nuevas y mejores soluciones a problemas relevantes. La Minería de Datos ha avanzado grandemente y en la actualidad tiene aplicaciones específicas a problemáticas concretas como la EDM, siglas en inglés para Minería de Datos educativa (Educational Data Mining), la EDM es una disciplina en evolución, su objetivo es la búsqueda, análisis y extracción de conocimiento dentro de un contexto educativo, para ello emplea las estrategias usadas en Minería de Datos sobre datos

educacionales, en función de resolver problemas que mejoren tanto el proceso enseñanza - aprendizaje (Ballesteros, 2013), y el apoyo a la toma de decisiones a las autoridades de los centros educativos (Moscoso-Zea, 2016).

La Universidad de Cuenca una Institución de Educación Superior Ecuatoriana actualmente dentro de la categoría A, persigue como parte de sus objetivos estratégicos, ser una Institución de Educación Superior de excelencia, con tal reconocimiento tanto a nivel nacional como internacional (Universidad de Cuenca, Plan Estratégico, 2010). Este término “de excelencia” es sinónimo de calidad integral de una institución, por tanto se refiere a que la Universidad de Cuenca, consolidará dicha excelencia en función de la mejora continua que todas sus áreas desarrollen. Si tomamos en cuenta el Modelo CONEA de Evaluación de desempeño (CONEA, 2009), el cual considera a las Instituciones de Educación Superior como un proyecto académico estructurado en cuatro dimensiones básicas: Academia, Estudiantes y entorno, Investigación y finalmente Gestión, al hablar de excelencia universitaria se habla de optimización y mejora en estas cuatro dimensiones. Ante este reto, cualquier esfuerzo por optimización es un valioso aporte hacia la excelencia universitaria, por tanto la aplicación de técnicas de Minería de Datos sería un valioso aporte pues a través de ella se puede comprender de mejor manera cómo se estructuran las variables que afectan positiva o negativamente problemas que afectan a la calidad brindada por la institución, incluso permitiendo crear modelos con los cuales identificar tempranamente situaciones que permitan a la universidad tomar medidas preventivas en lugar de correctivas.

La Universidad de Cuenca al contar con un gran volumen de datos albergados en sus sistemas de información, los que han sido recolectados a lo largo de las últimas décadas, han servido principalmente con fines transaccionales. Con el objeto de explotar esa información que hasta el momento se encontraba aislada en diferentes bases de datos, la DTIC se encuentra implementando un sistema de Data Warehouse, a través del cual poder visualizar y comprender de una manera rápida y sencilla la información que reflejan el funcionamiento de la institución, y con ello mejorar el proceso de toma de decisiones. Sin embargo el aprovechamiento de los datos va más allá de su visualización en cubos, tableros, y la construcción de indicadores (Han, J., Kamber, M., 2006); el siguiente paso en este camino, es el análisis profundo de dichos datos con el fin de sacar a la luz patrones ocultos, sustentando relaciones que a simple vista no existen, y de esta manera generar nuevo conocimiento, el cual nos permita afrontar los problemas de una manera diferente (Yoandry Pacheco, 2015).

2 MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Minería de Datos

Para hablar de Minería de Datos es necesario mencionar al término KDD - Knowledge Database Discovery, o Descubrimiento de Conocimiento en Bases de Datos, KDD fue concebido en 1989 para destacar que ***el conocimiento*** es el producto final del descubrimiento basado en los datos (Fayyad, 1996); en este mismo trabajo el autor sugiere que la Minería de Datos es una de las etapas dentro del proceso KDD, que mientras la Minería de Datos consiste en la aplicación de algoritmos específicos para la extracción de patrones, KDD se enfoca en todo el proceso de descubrimiento de conocimiento, incluyendo las fases de: selección, pre procesamiento, transformación, Minería de Datos e interpretación / evaluación, tal como se muestra en la Figura 1.

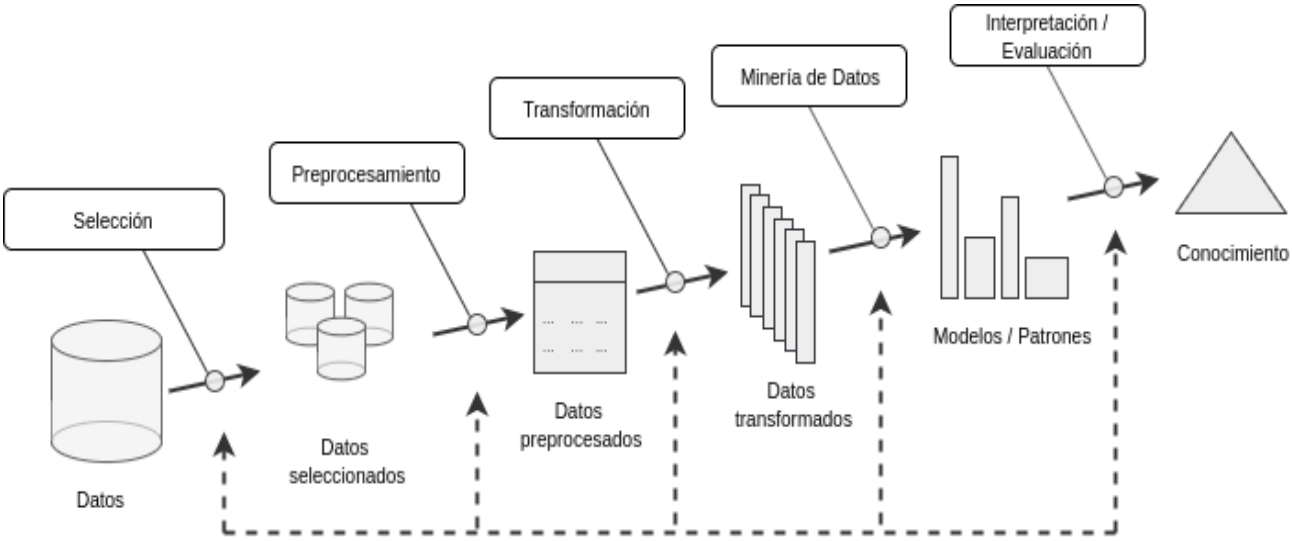


Figura 1. Pasos que componen el proceso KDD (Fayyad, 1996)

El objetivo final de KDD es encontrar patrones novedosos y útiles dentro de un conjunto de datos, y que estos patrones posibiliten el descubrimiento de un conocimiento igualmente válido y relevante aplicado a un determinado problema; por tanto KDD se encarga también de gestionar aspectos que van desde el almacenamiento y recuperación de los datos, hasta la interpretación de los resultados de la Minería de Datos tras lo cual se genera el conocimiento, a todo el proceso se lo organiza en los siguientes pasos: el paso de ***selección*** corresponde a la captura del conjunto de datos sobre los que el descubrimiento será hecho, el paso de ***preprocesamiento*** en cambio consiste en realizar operaciones de depuración y limpieza sobre el conjunto de datos con el fin de que sean consistentes, el paso de ***transformación*** consiste en la realización de operaciones de reducción de dimensionalidad, el paso de ***Minería de Datos*** consiste en la búsqueda de patrones sobre los datos, a través del empleo de técnicas estadísticas, machine learning, etc.,

el paso de ***interpretación / evaluación*** consiste en la interpretación y evaluación de los patrones extraídos. (Azevedo, 2008).

Si bien la mayoría de autores coinciden en que Minería de Datos corresponde a una etapa dentro del Proceso de Descubrimiento en los Datos, en la industria el término Minería de Datos ha sido más popular por lo que en varios artículos se los consideran sinónimos; incluso la metodología estándar de la industria CRISP-DM resalta este término incluyéndolo en su mismo nombre Cross Industry Standard Process for Data Mining. Esta metodología aparece en la década del 2000, con el objetivo de organizar el desarrollo de un proyecto de Minería de Datos en una serie de seis fases como muestra la Figura 2. CRISP-DM está descrita en términos de un modelo de proceso jerárquico consistente en un conjunto de tareas distribuidas en cuatro niveles de abstracción que van de lo general a lo específico, estos niveles son: fase, tarea genérica, tarea especializada, e instancia de procesos. Esta metodología como su nombre indica corresponde al estándar de la industria, y también a la metodología más utilizada desde su lanzamiento (Olson, 2008).

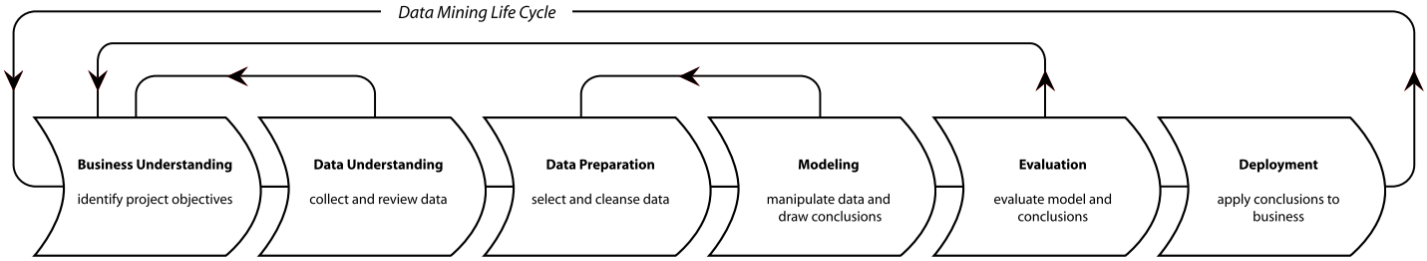


Figura 2. Una guía visual de la metodología CRISP - DM (Chapman *et al.*, 2000)

Si se compara el enfoque de KDD frente a CRISP-DM, el segundo propone extender el alcance de KDD incluyendo dos etapas que en el segundo no estaban identificadas: la de *Comprensión de Negocio* (Business Understanding) y la de *Despliegue* (Deployment), en la Tabla 1 se presenta un cuadro comparativo.

Tabla 1. Correspondencia entre KDD y CRISP - DM

KDD	CRISP - DM
------ Pre KDD	Comprensión del negocio
Selección	Comprensión de los datos
Pre procesamiento	
Transformación	Preparación de los datos
Minería de datos	Modelamiento
Interpretación / Evaluación	Evaluación
------ Post KDD	Despliegue

FUENTE: (Azevedo, 2008)
ELABORACIÓN: Gustavo Cordero

CRISP-DM al corresponder a una metodología más formal es más exhaustiva en su ejecución que KDD. Al basarse en un proceso de modelo jerárquico de gran alcance, se ajusta a las diversas realidades de los proyectos, además se preocupa en generar la suficiente documentación del proceso seguido con el objeto de que este sea replicable. Es decir, está diseñada para aplicarse de forma personalizada a diversos tipos de proyecto, de forma que el proceso seguido genere la suficiente documentación para que pueda servir de modelo replicable en otros proyectos similares. En cambio KDD es un proceso que se enfoca más hacia la experimentación, de cierta manera en KDD se conoce la pregunta a resolver y los datos disponibles, por ello se enfoca en ir directamente de los datos disponibles hacia la experimentación.

2.2 Minería de Datos Educativa (EDM) y Analítica de Aprendizaje (LA)

EDM (Educational Data Mining) ha sido una disciplina en evolución, desde 2009 ha contado con una comunidad de investigadores que ha crecido y se ha desarrollado, esto ha permitido incluso que desde ese mismo año se lleve a cabo una serie de conferencias anuales denominadas EDM (http://www.educationaldatamining.org/IEDMS/events), constituyendo también su propio Journal JEDM (Journal of Educational Data Mining), posteriormente la publicación del libro Handbook of Educational Data Mining de Chapman & Hall/CRC, es otro hito importante que ha dado impulso a esta disciplina. (Baker, 2009). Por otro lado la ***Analítica de Aprendizaje*** se ha definido como un área de investigación y aplicación que está relacionada con el análisis académico, el análisis de acciones y el análisis predictivo. La analítica del aprendizaje se basa en una gama más amplia de disciplinas académicas que la Minería de Datos Educativa, incorporando conceptos y técnicas de la ciencia de la información y la sociología, además de la informática, la estadística, la psicología y las ciencias del aprendizaje.

Como se mencionó en el capítulo introductorio, EDM corresponde a la aplicación de técnicas propias de Minería de Datos hacia el ámbito educativo. EDM es un paradigma orientado a la generalización de modelos, métodos y algoritmos para la exploración de datos propios de un contexto educativo, con el fin de encontrar y analizar patrones que caracterizan el comportamiento de los alumnos en base a sus logros, evaluaciones y el dominio de conocimiento alcanzado con los diversos mecanismos de aprendizaje - enseñanza que hoy en día están disponibles en las instituciones educativas, sobre todo de nivel superior. La Minería de Datos Educativa busca aportar en la generación de modelos educativos en los cuales se fomenten nuevas técnicas y herramientas que mejoren el nivel participativo de los estudiantes, (Ballesteros, 2013).

Un aspecto importante a destacar acerca de los métodos de EDM, es que a menudo difieren de los estándares aplicados en Minería de Datos general, esto debido a que

en EDM existe la necesidad de presentar explícitamente la *jerarquía multinivel* y la *dependencia* entre los datos educativos. Por esta razón, cada vez es más común ver el uso de modelos extraídos de la literatura psicométrica en publicaciones de Minería de Datos Educativos (Baker, 2010).

2.3 Métodos de EDM

Las metodologías usadas en EDM provienen de varias fuentes, pero las dos más importantes han sido por un lado *Minería de Datos y Analítica*, y por otro lado de *Psicometría y Medición Educativa*. En muchos casos, las características específicas de los datos educativos han dado lugar a diferentes métodos que han desempeñado un papel más significativo en EDM que en la Minería de Datos en general, o han dado lugar a adaptaciones a los métodos psicométricos existentes. A continuación se detallarán algunas de los principales métodos aplicados en EDM (Moscoso-Zea, 2016; Bhagoriya, 2017).

Métodos de Predicción: en predicción el objetivo es desarrollar un modelo que permita inferir aspectos simples de los datos (*variable predicha*), en base a la combinación de otros aspectos de los datos (variables predictoras). Para desarrollar un modelo de predicción es necesario conocer los valores de la variable predicha para un grupo de datos pequeño, comúnmente denominado grupo de entrenamiento, el modelo entonces es generado para dicho grupo, posteriormente es validado con otro grupo de datos, el denominado grupo de prueba, con dicha validación y en base a la precisión alcanzada, se podrá asegurar si el modelo puede emplearse o no a gran escala. Los tres modelos de predicción más comunes en EDM son: clasificadores (classifiers), regresores (regressors) y estimación de conocimiento latente (latent knowledge estimation). En clasificadores la variable predicha puede ser binaria o categórica, algunos métodos populares son: árboles de decisión, bosques aleatorios, reglas de decisión, regresión gradual y regresión logística. En regresión, la variable predicha es una variable continua (numérica), el regresor más popular en EDM es regresión lineal. En estimación del conocimiento latente, el cual realmente corresponde a un tipo especial de clasificador, se evalúa el conocimiento de un estudiante sobre determinados aspectos y conceptos específicos, en base a sus patrones de correctitud en esas habilidades. Los modelos utilizados en el aprendizaje en línea suelen diferir de los modelos psicométricos utilizados en las pruebas de papel o en las pruebas adaptativas por computadora, porque con una aplicación de aprendizaje interactivo, el conocimiento del estudiante está cambiando continuamente.

Descubrimiento de estructuras: el objetivo es descubrir estructuras dentro de los datos sin tener una idea a priori de que es lo que se encontrará, algo muy diferente al objetivo en predicción, donde había una variable específica que el investigador trata de modelar, en cambio en descubrimiento de estructuras no hay una variable de

interés específica, se trata de determinar qué estructura se obtiene de forma natural de los datos. Los enfoques más comunes son los siguientes: clustering, análisis de factores, análisis de redes sociales y descubrimiento de estructuras de dominio. En clustering el objetivo es encontrar ***"puntos"*** en los datos que los agrupen naturalmente, formando los denominados ***"clusters"***, este tipo de algoritmos son útiles en casos donde las categorías más comunes dentro de los datos no son conocidos previamente, como por ejemplo encontrar agrupación de estudiantes en base a distintos criterios (Beal, 2006). Los modelos para análisis de redes sociales (SNA) son desarrollados a partir de la relación e interacción entre los actores individuales, así como de los patrones que emergen de su relación e interacción. Por ejemplo, para identificar proyectos grupales efectivos o no, se podría recurrir a analizar visualmente la fuerza de las conexiones del grupo. Los patrones de interacción y conectividad pueden indicar una mayor posibilidad de éxito académico, así como el sentido de compromiso del alumno con el curso. (Macfadyen, 2010). El descubrimiento de estructuras de dominio consiste en encontrar la estructura de conocimiento dentro de un dominio educativo. Por ejemplo, como contenidos específicos pueden mapearse a componentes específicos de conocimiento o habilidades, en los estudiantes. (Baker, 2014).

Minería de relaciones: el objetivo es descubrir relaciones entre variables dentro de un conjunto de datos que posee un gran número de variables; se podría aplicar para identificar cuáles son las variables que poseen la relación más fuerte con una determinada variable de interés, o se podría buscar descubrir cuál relación entre dos variables es la más fuerte. Existen cuatro tipos de minería de relaciones: reglas de asociación, minería de correlaciones, minería secuencial de patrones, minería causal de datos (Moscoso-Zea, 2016). En reglas de asociación el objetivo es encontrar reglas del tipo SI - ENTONCES, de modo que si cierto conjunto de valores en determinadas variables es encontrado, otra variables generalmente tendrá un valor específico. En minería de correlación el objetivo es encontrar correlaciones lineales positivas o negativas entre variables. En minería secuencial de patrones el objetivo es encontrar asociaciones temporales entre eventos. En minería causal el objetivo es encontrar si un evento es la causa de otro evento, por ejemplo encontrar que factores afectan un pobre rendimiento de los estudiantes en la clase (Fancsali, 2012).

Descubrimiento con modelos: el modelo construido en base a un análisis de Minería de Datos, es aplicado sobre datos para evaluar el fenómeno que el modelo identifica, la predicción entonces es utilizada como entrada en otro método de Minería de Datos. Por ejemplo utilizar los resultados de un modelo predictivo como variable predictora en un segundo modelo de predicción y esto en múltiples niveles (Baker, 2014).

2.4 Estado del Arte

Si bien la Minería de Datos Educacionales (EDM) apareció antes que la Analítica de Aprendizaje (LA), ocupando en los primeros años a los investigadores en la aplicación y especialización de varias técnicas propias de DM al contexto educacional, en los últimos años ha destacado el interés por el desarrollo de aplicaciones más enfocadas al Learning Analytics, es decir emplear las técnicas ya definidas en EDM pero a contextos más relacionados a mejorar la experiencia de aprendizaje y su efectividad. Podríamos decir que la EDM ha dado las pautas y técnicas base que actualmente a través de LA se aplican a contextos híbridos con otras temáticas como: psicología, docencia, sociología, ciencias del aprendizaje, etc. Por tanto en la presente sección procederemos a indicar un estado del arte sobre la aplicación de estas técnicas a problemas educacionales, más que diferenciar si estos corresponden a EDM o LA.

En la Tabla 2 se presenta una análisis completo sobre varios problemas del contexto educativo que son apoyados a través de técnicas de Minería de Datos (Romero & Ventura, 2010).

Tabla 2. Principales aplicaciones de EDM

Categorías de estudios de EDM	Descripción	Técnicas empleadas
Análisis y Visualización de Datos	Información para toma de decisiones y retroalimentación a los docentes.	• Estadística • Técnicas de visualización de información. • Reglas de asociación • Clustering • Clasificación
Retroalimentación para apoyo a docentes		
Recomendaciones para estudiantes	Dar recomendaciones directamente a los estudiantes, actividades a realizar, enlaces a visitar, etc.	• Minería de reglas de asociación • Clustering • Minería de patrones secuenciales.
Predicción del rendimiento estudiantil	Descubrir características de los estudiantes. Crear modelos cognitivos de los estudiantes. Detección de estudiantes que poseen problemas o comportamiento inusual. Clasificación en grupos de alumnos de acuerdo a sus características personales.	• Regresión • Clasificación
Modelamiento de estudiantes		• Clustering • Clasificación
Detección de comportamientos indeseables en los estudiantes		
Agrupamiento de estudiantes		
Análisis de redes sociales (SNA)	Analizar las relaciones entre estudiantes	• Filtrado colaborativo
Desarrollo de mapas de concepto	Ayuda al docente para la construcción de mapas conceptuales	• Reglas de asociación • Minería de texto

Construcción de material didáctico	Ayudar al docente en la construcción de material didáctico y contenido de aprendizaje	• Clustering • Hybrid Unsupervised DM
Planificación y programación	Mejorar el proceso tradicional a través de la planificación de futuros cursos, apoyar al estudiante con su planificación, etc.	• Clasificación • Prediction • Clustering • Visualización de información.

FUENTE: (Romero & Ventura, 2010)
ELABORACIÓN: Gustavo Cordero

Otra categorización es la propuesta en la revisión de 300 artículos presentada por (Moscoso & Zea, 2016), ahí se observa que las principales áreas de experimentación han sido las siguientes:

1. Predicción de evaluaciones finales o rendimiento de los estudiantes,
2. Predecir el perfil y comportamiento de aprendizaje de los estudiantes,
3. Mejorar el soporte del docente,
4. Mejorar la gestión de la colaboración en ambientes educacionales, etc.

Siendo la primera área de aplicación, la que predomina por el mayor número de experimentos realizados. Del análisis realizado, el autor indica que la mayor cantidad de experimentos en EDM han empleado técnicas de clasificación, asociación y predicción para resolver sus casos de estudio. Siendo CRISP-DM el framework más utilizado en los proceso de EDM.

Por otro lado existe evidencia de la efectividad alcanzada con la aplicación de analítica de aprendizaje en diferentes iniciativas (Sclater & Mullan, 2017):

- Modelos predictivos precisos en la identificación de estudiantes con riesgo: las instituciones educativas pueden dar un seguimiento individual al compromiso, consecución y progreso de los estudiantes.
- Intervenciones institucionales efectivas: sólo cuando las acciones son realizadas tomando como base al estudiante es cuando el valor de la analítica de aprendizaje (LA) se hace claro.
- Cambios en la conducta de los estudiantes: por ejemplo hay métricas que los estudiantes que han empleado la herramienta informática para comprar su actividad en los Sistemas Virtuales de Aprendizaje con otros estudiantes tuvieron 1.92 más probabilidades de pasar con una calificación C o superior.

3 OBJETIVOS

3.1 Objetivo General

Establecer un proceso para el Descubrimiento de Conocimiento en el sistema de Data Warehouse de la Universidad de Cuenca mediante la aplicación técnicas de Data

Mining, y su aplicación sobre tres ámbitos concretos: el éxito académico, evaluación docente y realidad socioeconómica.

3.2 Objetivos Específicos

1. Identificar qué criterios entre asignaturas, calificaciones y datos de la ficha socioeconómica son los más relacionados al éxito académico en alumnos nuevos y analizar si existe diferencia dependiendo de la facultad.
2. Determinar la factibilidad de construir una herramienta predictora de la deserción estudiantil en nuevos alumnos.
3. Evaluar la relación existente entre el proceso de evaluación docente y el éxito académico de los alumnos, con la obtención de una clasificación natural de los resultados en las evaluaciones docentes.
4. Identificar oportunidades de información que potencien la aplicación de la Minería de Datos para futuros estudios.

El objetivo general se lo desarrollará en la sección 4, los objetivos específicos del uno al tres se desarrollan en el capítulo 4, el objetivo específico 4 se lo desarrolla dentro de la sección Trabajos futuros del Capítulo 6.

3.3 Alcance

El establecimiento de un proceso se refiere a la especificación de un modelo BPMN genérico para la aplicación de Minería de Datos en la Universidad de Cuenca, dicho proceso será compatible con las metodologías formales como CRISP DM a la vez que se buscará dar algún enfoque de metodología ágil.

Tanto para el modelamiento de dicho proceso como para su ejecución, se tomará como principal fuente de datos el sistema de Data Warehouse de la Universidad, puesto que el esfuerzo de consolidación de la información de las fuentes primarias de los datos debería ser uno sólo, de manera que se aproveche la información ya centralizada en el sistema de Data Warehouse.

Las técnicas de minería de datos a emplear serán las más utilizadas en Minería de Datos Educacional o EDM (Cristobal, R. y otros, 2000): Clasificación, Clustering, Reglas de Asociación y Predicción (Moscoso-Zea, Oswaldo, 2016).

4 PROCESO GENÉRICO PARA LA EJECUCIÓN DE PROYECTOS DE MINERÍA DE DATOS EN LA UNIVERSIDAD DE CUENCA

Al ser el objetivo del presente capítulo el sugerir un proceso estándar para la aplicación de proyectos de Minería de Datos dentro de una Universidad, es natural que por su específico giro de negocio esté interesada en la ejecución de proyectos dentro del contexto de EDM; sin embargo en este caso la Universidad de Cuenca es un centro académico grande, que atiende a más de 16.000 alumnos, por lo que para su adecuada gestión ha implementado varias áreas y procesos complementarios a los académicos como: planificación, contabilidad, tesorería, inventario, talento humano, legal, entre otros. Áreas que generan volúmenes de datos considerables, cuyo análisis generará un valioso aporte, pero que cae en el contexto general de Minería de Datos y no en el contexto específico de EDM. Por esta razón, se hace necesario que el proceso a diseñar sea aplicable en ambos contextos.

Para fundamentar adecuadamente las bases metodológicas del proceso a diseñar, a continuación se presentan brevemente algunas propuestas realizadas por varios autores, las cuales han servido como base para construir el proceso genérico sugerido en este trabajo.

4.1 CRISP DM

Como se indicó anteriormente, la metodología CRISP-DM corresponde al estándar de la industria, el cual está descrito en términos de un modelo de proceso jerárquico consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos; y organiza el desarrollo de un proyecto de Minería de Datos, en una serie de seis fases como se muestra en la Figura 3.

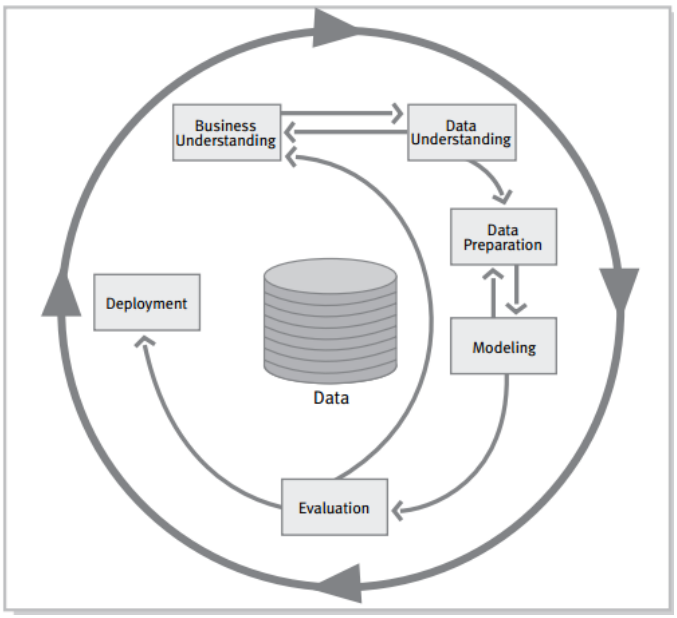


Figura 3. Modelo CRISP - DM

1. Comprensión del negocio

2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Implantación

4.2 ASD - DM

Es un framework denominado ASD - DM propuesto en Alnoukari et al, 2008, este propone la combinación de las características propias del Desarrollo de Software Adaptativo o ASD, con los pasos generales utilizados en proyectos de Minería de Datos. La propuesta de los autores consisten en la estructuración de las etapas especificadas por CRISP-DM en las tres fases propias de ASD: Especulación, Colaboración y Aprendizaje, como se muestra en la figura 4. Estas fases permiten que la gestión del proyecto posea los siguientes fundamentos: orientación a las características, iteratividad, gestión de riesgos y tolerancia a los cambios.

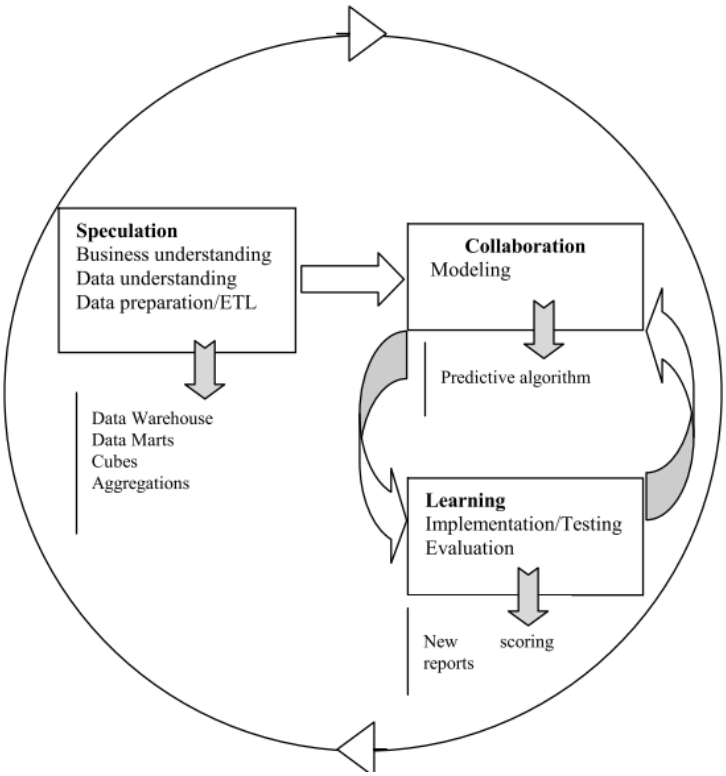


Figura 4. Framework para Minería de Datos Predictiva basado en la desarrollo de software adaptativo ASD

Lo interesante de esta propuesta es el componente ágil que se busca insertar en lugar del ciclo tradicional Planifica - Diseña - Construye, que puede suponer CRISP-DM. Y que tiene mucha coherencia puesto que normalmente un proyecto de Minería de Datos, posee un factor importante de incertidumbre sobre los datos disponibles, su calidad, la precisión de los modelos, etc., aspectos que solamente con la realización

del experimento (Modelamiento) se podrían verificar. Ante esta perspectiva la fase de planificación termina siendo menos útil mientras más incertidumbre exista en el proyecto. Probablemente lo más acertado sería, una vez definido el problema o pregunta, realizar un proceso liviano sobre las primeras fases del proceso (comprensión del negocio), y llegar rápidamente hasta la experimentación (comprensión de los datos y modelamiento), esto permitiría al analista involucrarse en el contexto del proyecto, los recursos disponibles y los resultados probables. Luego de lo cual ejecutar una nueva iteración sobre todas las fases del proceso, y en ella definir los puntos previamente pasados por alto o poco desarrollados, como sería caso de la planificación del proyecto.

4.3 Modelo de Proceso KDDA en caracol

Una propuesta similar a lo planteado al final del punto anterior es la abordada por Li Yan en su tesis doctoral “New Artifacts For The Knowledge Discovery Via Data Analytics (Kdda) Process”. El autor en el capítulo 4 presenta un modelo denominado “Modelo de Proceso KDDA en caracol” (A Snail Shell KDDA Process Model) concordante con las fases y tareas de CRISP-DM, sin embargo sus dos principales aportes corresponden a la marcada importancia que se le da a la primera fase del proyecto Formulación del Problema, y posteriormente al carácter iterativo de las siguientes fases, como se puede apreciar en la Figura 5, ahí se hace explícita la posibilidad de que después de cada fase se puede nuevamente regresar a las fases iniciales para corregir o precisar algo en el proceso. Vale la pena indicar que el mismo CRISP-DM considera igualmente esta necesidad, y deja abierta la posibilidad de moverse entre las distintas fases y actividades en base a las particularidades de cada proyecto.

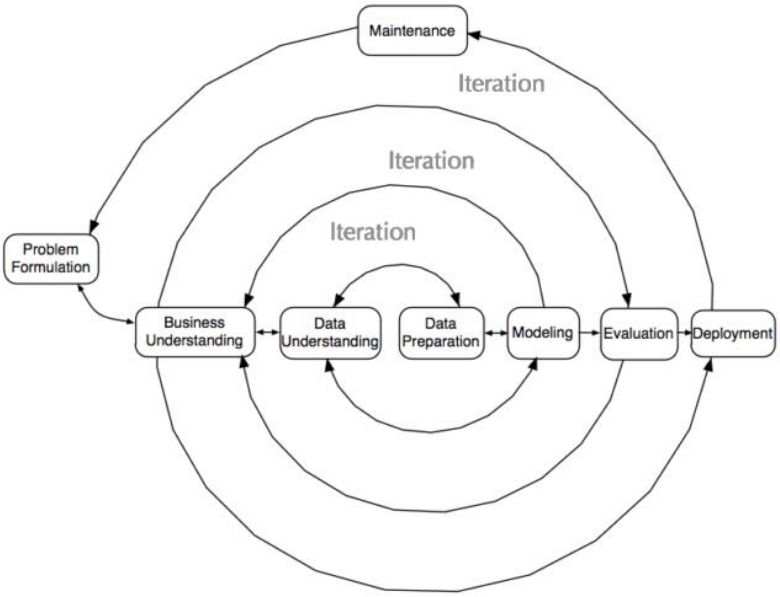


Figura 5. Modelo de Proceso KDDA en caracol

4.4 SEMMA

Una metodología creada por el instituto SAS (Statistical Analysis System), el nombre de la metodología corresponde al acrónimo de Sample (extracción), Explore (exploración), Modify (manipulación), Model (modelado), Assess (valoración), como se muestra en la Figura 6. SEMMA se define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocido (Fischer, 2012).

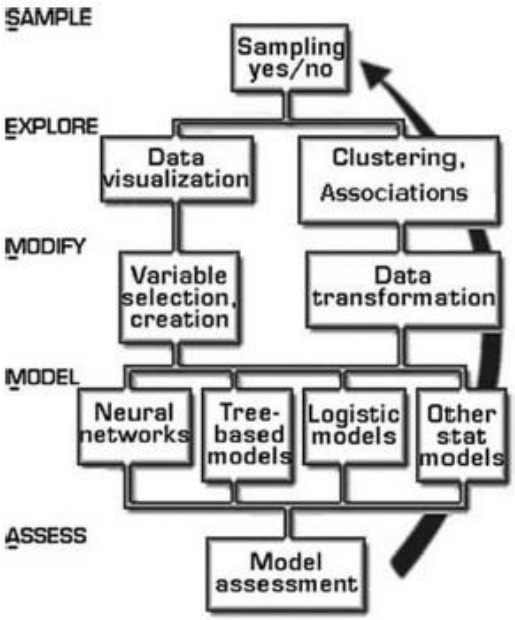


Figura 6. Modelo SEMMA

Esta metodología se enfoca principalmente en las fases de Modelado, y no abarca aspectos específicos del negocio, por lo que posee un alcance menor al que propone CRISP-DM y es más experimental.

4.5 Fundamentos del proceso

En base a la revisión realizada sobre los modelos existentes, es claro que CRISP-DM es el modelo de mayor alcance y al que tratan de acoplarse los demás, presentando además las siguientes ventajas: a) corresponde al estándar de la industria, b) es actualmente la metodología más utilizada a nivel de proyectos de Minería de Datos (Moscoso-Zea, 2016), c) es flexible y versátil, su utilización no implica una camisa de fuerzas, d) puede mapearse de un esquema genérico a un modelo especializado, e) abarca etapas anteriores y posteriores que otras metodologías no lo hacen, lo que le da un alcance integral desde una perspectiva de negocio (Azevedo, 2008).

De las metodologías analizadas se puede decir que todas guardan coherencia con CRISP-DM. Puntualmente *"ASD-DM"* y *"Modelo de Proceso KDDA en caracol"* buscan dotar de agilidad al proceso metodológico, lo cual resulta atractivo dado de que al iniciar un proyecto de Minería de Datos, el factor de incertidumbre es generalmente alto, puesto que se trata justamente proyectos de tipo experimental y de descubrimiento. Por tanto, ejecutar actividades demasiado profundas de planificación y análisis durante las fases iniciales de compresión de negocio, comprensión de los datos y preparación de los datos, podrían representar un tiempo

considerable hasta verificar los datos disponibles y la precisión alcanzable de los modelos; mientras que desarrollar una primera iteración rápidamente con la intención de realizar un experimento, permitiría disponer de una prueba de factibilidad del proyecto y de los resultado a alcanzar, y a la vez dotaría al analista de mayores detalles del contexto del proyecto, los recursos disponibles, etc. En las siguientes iteraciones en cambio el objetivo si sería profundizar en las fases, tareas y salidas sugeridas por CRISP-DM, siempre que dichas actividades generen un aporte y sean aplicables al contexto del proyecto, con el objeto de asegurar su calidad, replicabilidad en caso de ser necesario y dado que la Universidad Cuenca cuenta con un Sistema de Data Warehouse, analizar si existen nuevas fuentes de datos que convengan incluir como Data Marts adicionales.

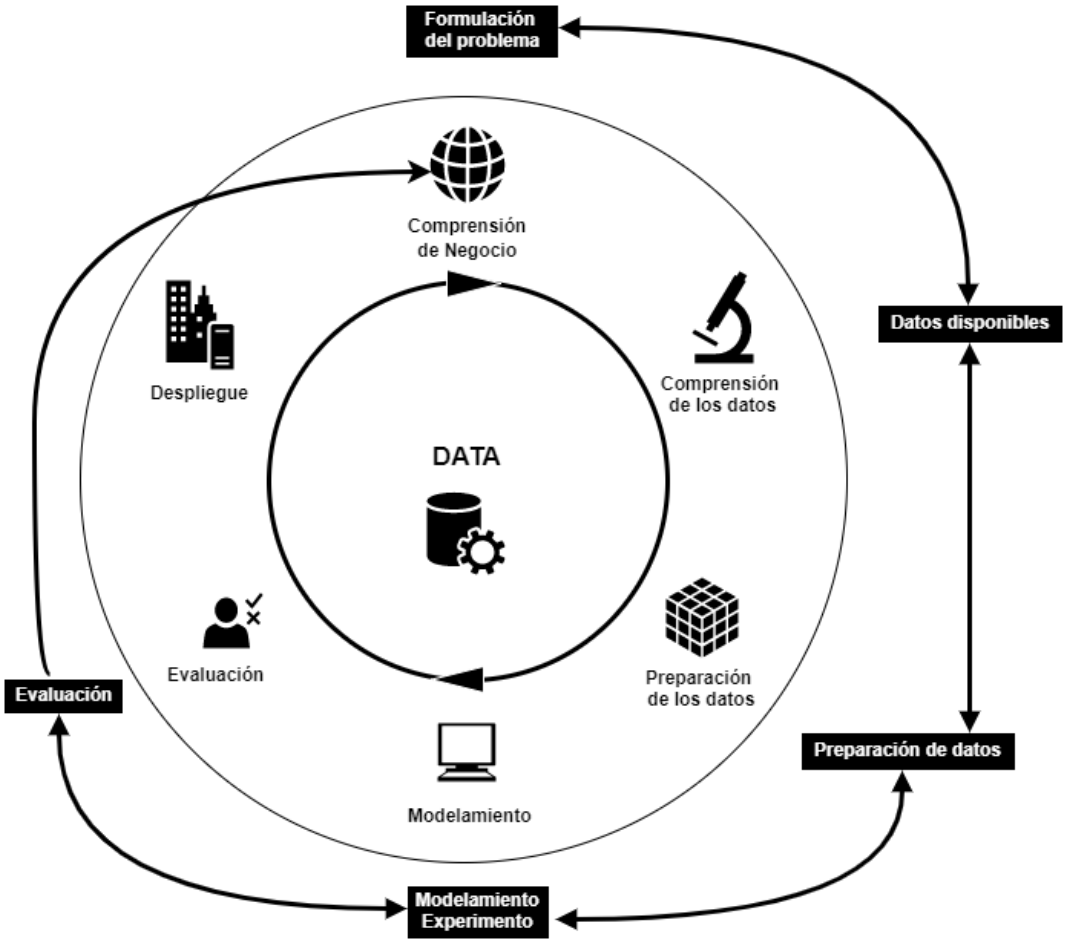


Figura 7. Ejecución en espiral sobre el Modelo CRISP-DM

Como se muestra en la Figura 7, CRISP-DM es el marco metodológico base del proceso, sin embargo con la intención de dotar de agilidad a los proyectos se ha resaltado una iteración inicial que parte de la formulación del problema (siguiendo la propuesta del Modelo de Proceso KDDA en caracol), tras lo cual inmediatamente se ejecuta un proceso rápido de experimentación a través de las fases de comprensión de los datos, preparación de los datos, modelamiento y evaluación, esto dará un mayor conocimiento del contexto del proyecto al analista y permitirá que la siguiente

iteración sea realizada con mayor especificaciones, el plan de proyecto sea más preciso, y la documentación generada sea adecuada para la replicación.

4.6 Descripción del proceso

En el presente punto se desarrolla la descripción de los procesos sugeridos, para el adecuado desarrollo y control de proyectos relacionados al descubrimiento de conocimiento en bases de datos de la Universidad de Cuenca, empleando técnicas de Minería de Datos.

Para la descripción de dichos procesos se ha empleado la notación BPMN (Modelo y Notación de Procesos de Negocio), la cual es una notación gráfica estandarizada que permite el modelado de procesos de negocio en un formato de flujo de trabajo (workflow) (Business Process Model and Notation, s.f). Esta notación tiene varias ventajas (Allweyer, 2013):

- Es un estándar internacional de modelado de procesos aceptado por la comunidad.
- Es independiente de cualquier metodología de modelado de procesos.
- Crea un puente estandarizado para disminuir la brecha entre los procesos de negocio y la implementación de estos.
- Permite modelar los procesos de una manera unificada y estandarizada permitiendo un entendimiento a todas las personas de una organización.

El proceso completo para su mayor comprensión se ha estructurado en tres niveles jerárquicos, los cuales se muestran en la Tabla 3.

Tabla3. Estructura jerárquica de los procesos

Proceso padre	Proceso central	Subprocesos
Iniciación de proyecto de Minería de Datos		
	Gestión macro de proyecto de Minería de Datos	
		Iteración de experimentación sobre CRISP-DM
		Iteración detallada sobre CRISP-DM

ELABORACIÓN: Gustavo Cordero

Cabe indicar que tanto el proceso de ***Iteración de experimentación sobre CRISP-DM*** como el de ***Iteración detallada sobre CRISP-DM***, son procesos que se levantan sobre el mismo marco metodológico CRISP-DM, la diferencia consiste en que el primero pasa por alto varias actividades que son realizadas en el segundo, por tanto ambos procesos son descritos en una misma sección más adelante.

Para el modelamiento de los procesos se ha empleado la herramienta Visual Paradigm en su versión 14.0.

4.6.1 Iniciación de un proyecto de Minería de Datos

Este proceso corresponde a la definición de las actividades necesarias que permitan al departamento de la DTIC identificar si para cumplir con un requerimiento de información realizada por un Stakeholder, se debe dar inicio a un proyecto de Minería de Datos.

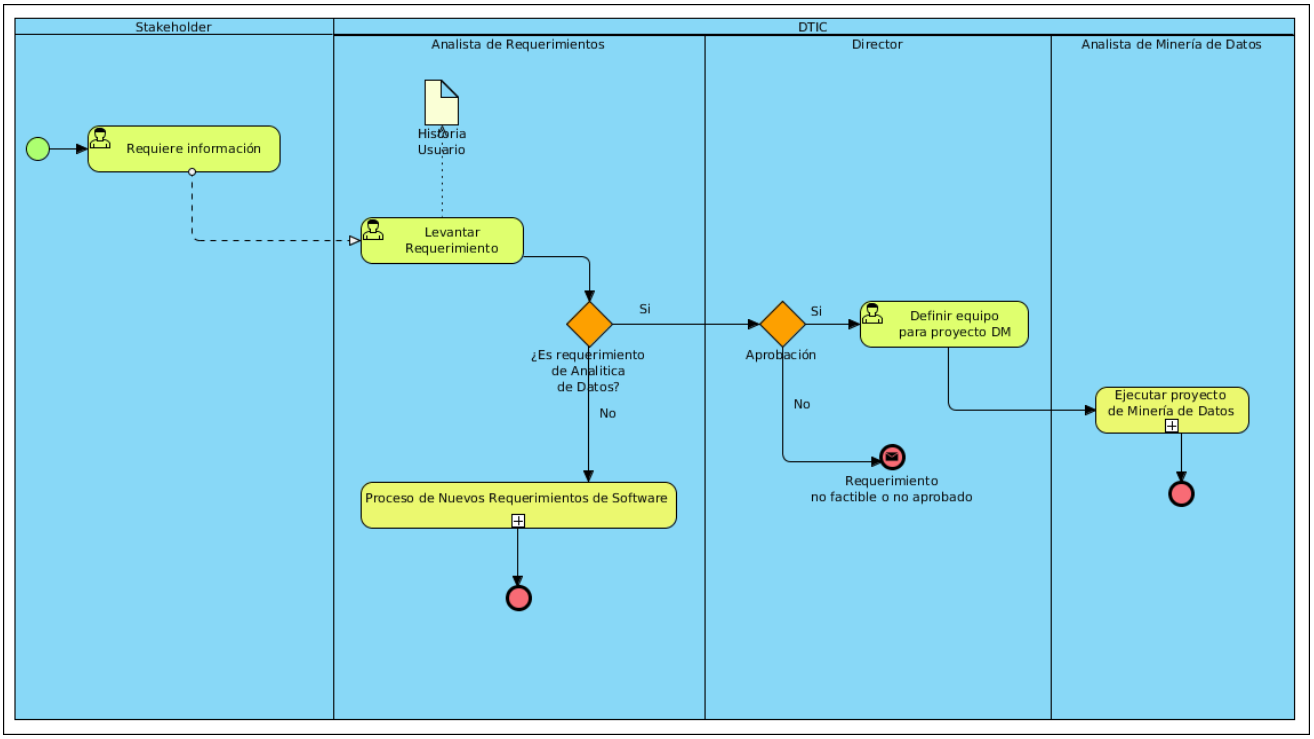


Figura 8. Iniciación de un proyecto de Minería de Datos.

A continuación se indican las características o perfiles de los actores involucrados en el proceso:

- **DTIC - Dirección de Tecnologías de la Información y Comunicación:** es el órgano encargado de la gestión, coordinación y ejecución de proyectos en el ámbito de las tecnologías de información y comunicación, orientados al mejoramiento de la calidad académica y administrativa de la Universidad. (Universidad de Cuenca, www.ucuenca.edu.ec).
- **Stakeholder:** corresponde a un funcionario de alguna de las dependencias de la Universidad quien presenta un requerimiento de información hacia la DTIC.
- **Analista de requerimientos:** especialista técnico encargado de la elicitación de requerimientos relacionados a TIC's.
- **Director - DTIC:** responsable de la aprobación del inicio de un proyecto interno dentro del departamento DTIC, en este caso de un proyecto de Minería de Datos.

- **Analista de Minería de Datos:** especialista técnico encargado de desarrollar el proyecto de Minería de Datos, es el responsable de dicho proyecto.

La descripción del proceso y las actividades que este involucra se encuentran descritas en la Tabla 4. Actividades del proceso Identificación y creación de un proyecto de Minería de Datos.

Tabla 4. Actividades del proceso Identificación y creación de un proyecto de Minería de Datos

Actividad	Responsable	Descripción detallada de la actividad
Requiere información	Stakeholder	Indica una necesidad de información mediante cualquier canal de comunicación: verbal, correo electrónico, etc.
Levantar Requerimiento	Analista de requerimientos	Describe formalmente la necesidad de información del usuario. Nota: ● En caso de existir se deberá adjuntar documentos de ejemplo. ● En caso de ser factible indicar cuál es el origen de los datos a analizar.
Proceso de Nuevos Requerimientos de Software	Analista de requerimientos	Seguir proceso para el análisis, diseño e implementación de nuevos requisitos de software.
Definir equipo para proyecto de DM	Director DTIC	Estructurar un equipo de trabajo para desarrollar el proyecto que satisfaga el requerimiento planteado, el equipo debe contener los siguientes roles: analista de Minería de Datos, especialista en integración de datos, SME o experto del tema a analizar (podría ser el mismo Stakeholder), experto en la estructura de datos. Nota: ● Se ejecuta sólo si el requerimiento corresponde a analítica de datos y ha sido aprobado previamente por el director
Ejecutar el proyecto de DM	Analista de Minería de Datos	Subproceso donde se ejecuta el proyecto.

ELABORACIÓN: Gustavo Cordero

4.6.2 Gestión Macro de un proyecto de Minería de Datos

Este proceso corresponde a un subproceso de *Iniciación de proyecto de Minería de Datos,* en este se desarrollarán las actividades iniciales del proyecto por parte del equipo que lo ejecutará. En este se define las actividades de nivel macro que el Analista de Minería de Datos y el Director de la DTIC, realizarán para iniciar el proyecto, revisar la factibilidad del mismo y de ser afirmativo lo último, ejecutar el proyecto de manera detallada. Las actividades de este proceso son de nivel superior a las definidas en el marco metodológico CRISP-DM.

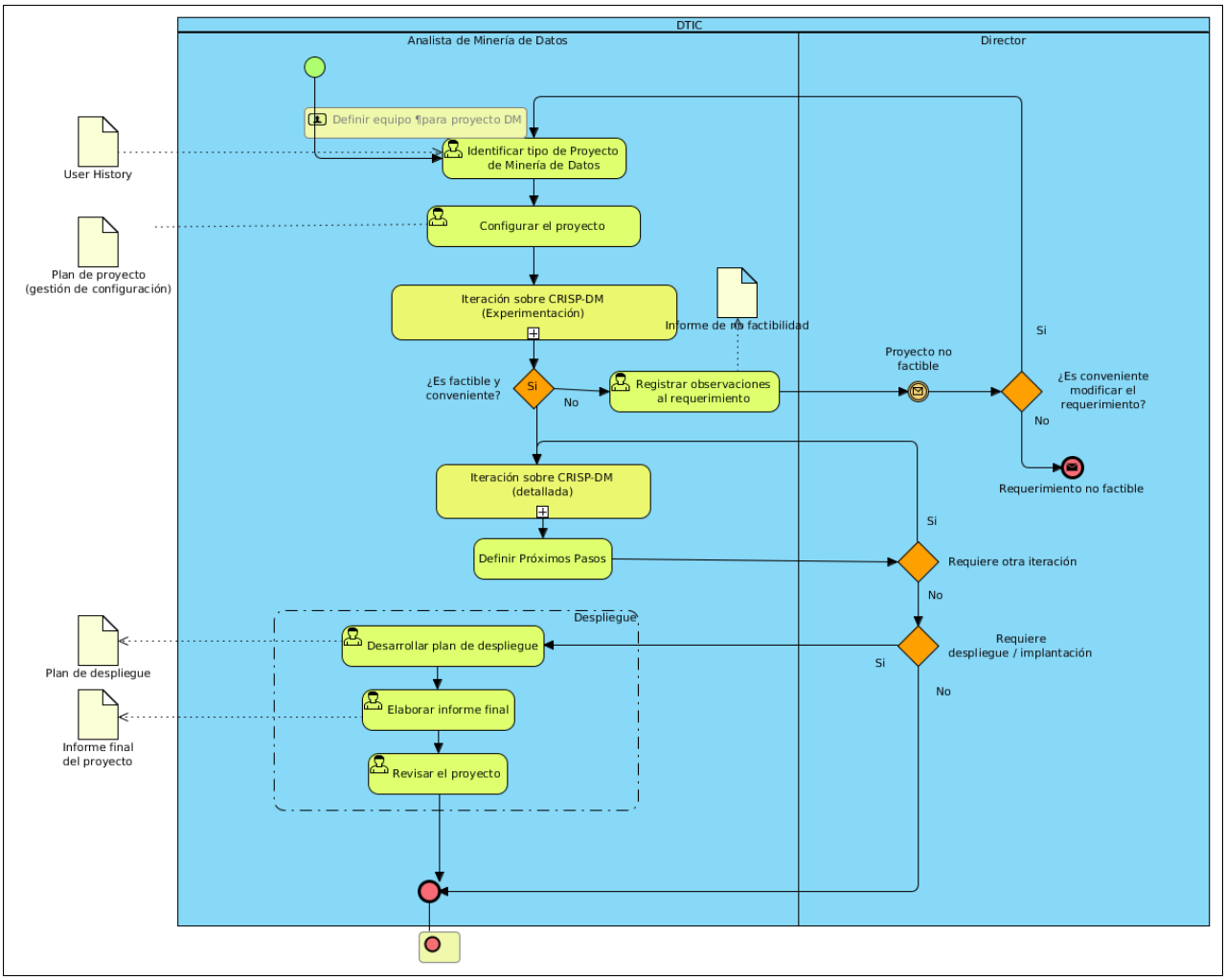


Figura 9. Gestión macro de proyecto de DM (Data Mining)

A continuación se indican las características o perfiles de los actores involucrados en el proceso:

- **Analista de Minería de Datos:** especialista técnico encargado de desarrollar el proyecto de Minería de Datos, es el responsable de dicho proyecto.
- **Director - DTIC:** máximo líder del departamento, persona que tiene la potestad de decidir si el proyecto puede cancelarse o no.

La descripción del proceso y las actividades que este involucra se encuentran descritas en la Tabla 5. Actividades de Gestión macro de proyecto de DM. Cabe indicar que aquellas actividades del proceso que son equivalentes a las definidas por CRISP-DM, pueden ser revisadas a mayor detalle por el lector en el ANEXO A. Traducción del modelo de referencia CRISP DM 1.0, para lo cual en cada caso se referencia la tarea específica.

Tabla 5. Actividades de Gestión macro de proyectos de DM

Actividad	Responsable	Descripción detallada de la actividad
Identificar tipo de proyecto	Analista de Minería de Datos	Analiza las características del requerimiento en base a levantamiento de requerimiento previamente realizado, pudiendo profundizar sobre el mismo mediante una nueva entrevista al stakeholder y/o al SME. El objetivo es definir características del proyecto como: extensión, complejidad, tipo de técnica a emplearse, etc. Nota: ● Como entrada se emplea el documento resultante de la elicitación de requerimiento: ERS o Historia de Usuario.
Configurar el proyecto	Analista de Minería de Datos	Consisten en la definición en un Plan de Configuración donde el analista seleccione las actividades del modelo metodológico que aplican al proyecto y los artefactos a emplear (dependiendo del contexto de DM), además de definir un repositorio para que dichos documentos sean adecuadamente accedidos y modificados por el equipo. Nota: ● Se puede aplicar aspectos generales de la Gestión de Configuración de CMMI (Modelo de Madurez de la Capacidad Integrado). ● Dependiendo del tamaño del proyecto el Plan de Configuración podría formar parte del Plan de Proyecto. ● Se debe diferenciar las actividades que serán seguidas en la iteración de experimentación a diferencia de las que se seguirán en la iteración detallada.
Iteración de experimentación sobre CRISP-DM (Subproceso)	Analista de Minería de Datos	Subproceso en el que se ejecutan las actividades de una primera iteración del proceso CRISP-DM para verificar la factibilidad del proyecto y mejorar la comprensión del contexto por parte del equipo.
Registrar observaciones al requerimiento	Analista de Minería de Datos	Describe las causas que justifican la no factibilidad o conveniencia de la ejecución del proyecto actual, para que el director del departamento pueda decidir sobre cancelar el proyecto o modificar el requerimiento primario. Nota: ● Se ejecuta si previamente se ha definido que el proyecto no es factible o conveniente (costo / beneficio)
Iteración detallada sobre CRISP-DM (Subproceso)	Analista de Minería de Datos	Subproceso en el que se ejecutan una segunda iteración de CRISP-DM, previamente se entiende verificada la factibilidad del proyecto, por lo que en este subproceso la intención es ser más detallado y recabar documentación para posibles futuras réplicas del proyecto. Nota: ● Se ejecuta una vez que se ha comprobado la factibilidad y conveniencia del proyecto.
Definir próximos pasos	Analista de Minería de Datos / Director de DTIC	Equivalente a la Tarea 5.3 de CRISP-DM Decidir cómo se va a proceder en base a los resultados de la evaluación, se decide en conjunto con el Director del Departamento si se debe ejecutar una nueva iteración, iniciar un nuevo proyecto o iniciar una fase de implementación / despliegue del proyecto.

Etapa 6: Despliegue		
Actividad	**Responsable**	**Descripción detallada de la actividad**
Desarrollar plan de despliegue, supervisión y mantenimiento.	Analista de Minería de Datos	Equivalente a las Tareas 6.1 y 6.2 de CRISP-DM Determinar una estrategia de despliegue dependiendo de los requerimientos iniciales, tipo de entregable: sistema, informe, conocimiento, respuesta, etc. Si los resultados son parte del negocio cotidiano y su ambiente, es necesario definir la supervisión y el mantenimiento del mismo. Salidas posible: ● Plan de despliegue. ● Plan de monitoreo y mantenimiento.
Elaborar informe final	Analista de Minería de Datos	Equivalente a la Tarea 6.3 de CRISP-DM Redactar un informe final, pudiendo ser sólo un resumen del proyecto, una presentación, etc. Salidas posible: ● Reporte final ● Presentación final
Revisar el proyecto	Analista de Minería de Datos	Equivalente a la Tarea 6.4 de CRISP-DM Evaluar lo correcto e incorrecto, lo que se hizo bien y lo que puede mejorarse. Salidas posible: ● Documentación de la experiencia.

ELABORACIÓN: Gustavo Cordero

4.6.3 Iteración de experimentación e Iteración detallada sobre CRISP-DM

Como se mencionó anteriormente ambos procesos se levantan sobre el mismo marco metodológico CRISP-DM, compartiendo características como los actores y artefactos empleados. Por ello en este punto se va a detallar un único proceso denominado ***Iteración sobre CRISP-DM***, y posteriormente se detallarán algunos aspectos a considerar correspondientes a la Iteración de experimentación.

A continuación se indican las características o perfiles de los actores involucrados en el proceso:

- **Analista de Minería de Datos:** especialista técnico encargado de coordinar el proyecto de Minería de Datos con los demás participantes y es el responsable de la ejecución de las actividades de análisis de datos.

- **DWH Experto:** especialista técnico experto en la implementación del sistema de Data Warehouse (DWH), por lo que debe conocer el dominio de los datos disponibles en los Data Marts que lo conforman. En caso de que existan orígenes de datos adicionales al DWH relevantes para el análisis, será la persona que tenga el criterio para decidir si estos orígenes son convenientes de ser incorporados al DWH y las políticas que definirán la implementación del nuevo Data Mart.
- **DBA - Experto en bases de datos:** especialista técnico, con conocimiento de las bases de datos y estructuras disponibles en los sistemas de la universidad, este rol podrá ser desempeñado por más de una persona dado el número de sistemas manejado por la universidad.
- **SME - Subject Matter Expert:** especialista experto en la temática de negocio sobre la cual el problema o requerimiento se desarrolla. Igualmente dada la extensión del problema planteado podría ser una o más personas.

Referenciando los lineamientos de CRISP-DM, el analista de Minería de Datos está en libertad de seleccionar qué actividades y artefactos va a emplear en su proyecto, tomando en consideración las características propias del mismo y el contexto de negocio en el cual se desarrolla, por tanto un proyecto de Minería de Datos que siga esta metodología no está obligado a ejecutar cada una de las actividades descritas en el presente proceso metodológico, de ello que las actividades listadas en la Tabla 6 son las sugeridas en base al estándar, y dependerá del criterio del Analista el desarrollo de las mismas. Dicha selección de las actividades a realizar y los artefactos a emplear debería ser realizada en la actividad ***Configurar el Proyecto*** dentro del proceso ***Gestión macro de proyecto de DM***, no obstante dicho artefacto podría cambiar durante la ejecución de las ***Iteraciones sobre CRISP-DM****.*

La Figura 10 corresponde a la primera parte del proceso, estos pasos involucran las fases iniciales de *Compresión del negocio* y *Comprensión de los datos*, en la primera el analista profundiza sus conocimientos sobre el negocio (partiendo ERS o Historia de Usuario), apoyado en el especialista en DWH y el DBA identifica los recursos de datos disponibles que apoyan el objetivo del proyecto, en esta fase se generan importantes artefactos como: objetivos de negocio, inventario de recursos disponibles, objetivos de Minería de Datos. Posteriormente en la fase de Comprensión de los datos, el analista profundiza en las características de los datos disponibles, igualmente se apoya en los especialistas en DWH y DBA para obtener los datos en un formato analizable al tiempo que se verifica que la calidad de los mismos sea adecuada.

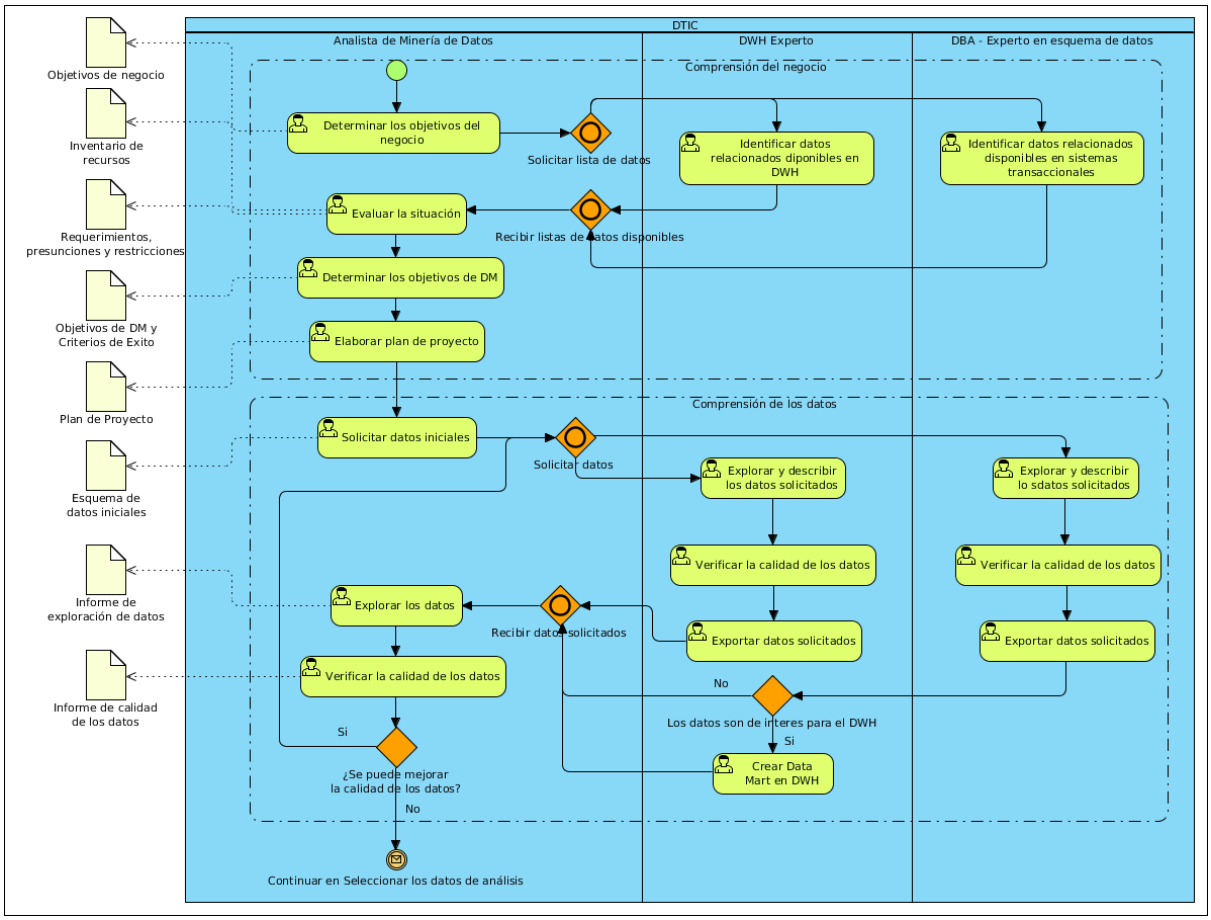


Figura 10. Iteración sobre CRISP-DM (Parte 1)

La Figura 11 corresponde a la segunda parte del proceso, la que corresponde a las fases de *Preparación de datos, Modelamiento* y *Evaluación*. En la fase de *Preparación de datos* el analista selecciona los datos que encuentra de aporte más significativo al análisis, los especialistas en DWH y el DBA, participan en la preparación y limpieza de los datos, al final de esta fase el analista cuenta con un Dataset adecuado para ejecutar el modelado.

La fase de *Modelado* en cambio es realizada por el Analista, en ésta fase se realiza la experimentación a través de la aplicación de una o más técnicas para el problema de Minería de Datos planteado. En la siguiente fase de *Evaluación*, los resultados obtenidos en el *Modelamiento* son analizados también por el SME dentro del contexto de Negocio, en este análisis se incluye también cualquier otro resultado producido a lo largo del proyecto.

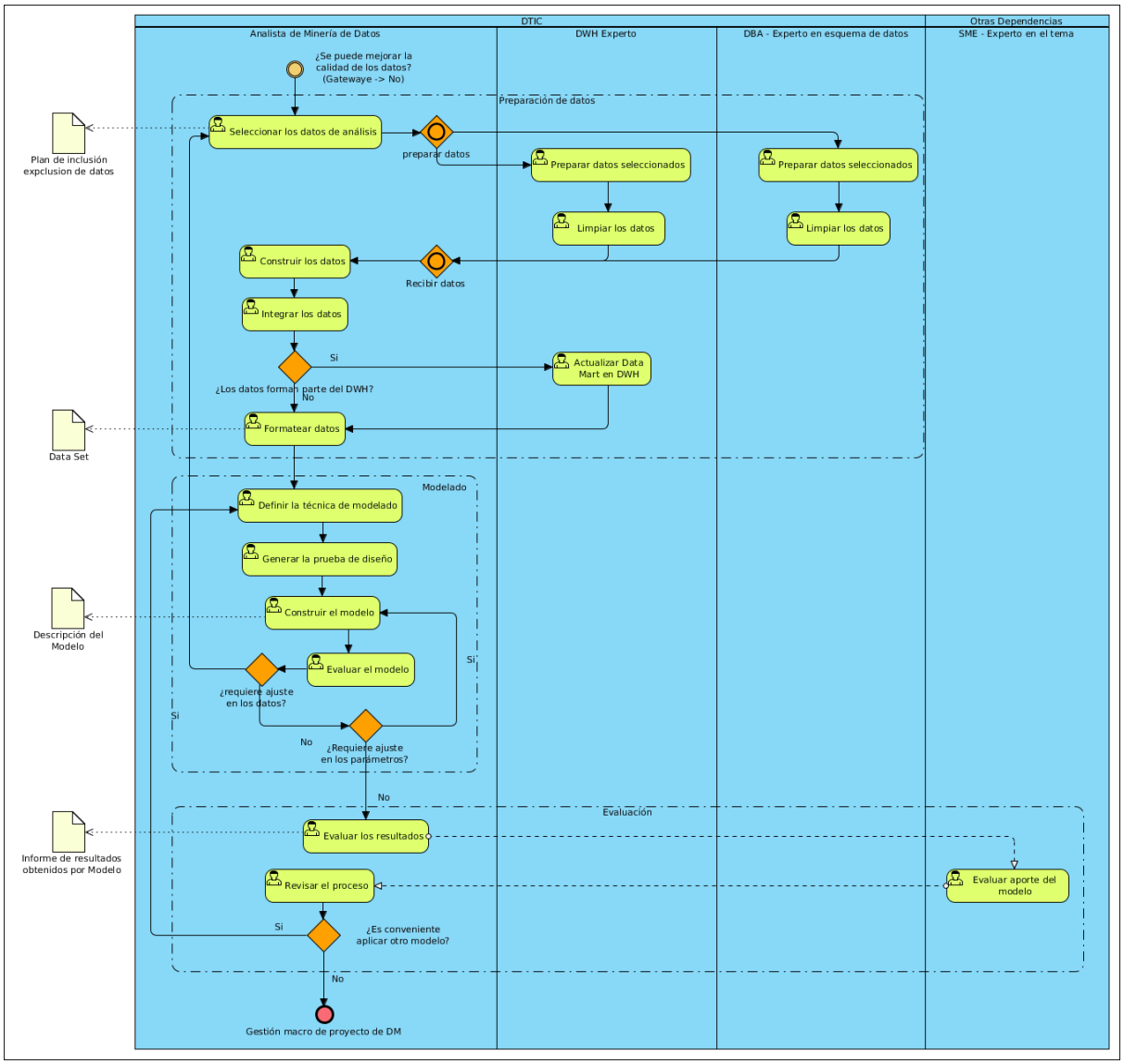


Figura 11. Iteración detallada sobre CRISP-DM (Parte 2)

En la Tabla 6. Actividades del proceso Iteración sobre CRISP-DM, se presenta un listado descriptivo de las actividades que componen este proceso, en esta tabla se han subrayado aquellas actividades se sugiere deben realizarse en la ***Iteración de experimentación.*** Cabe indicar también, que aquellas actividades del proceso que son equivalentes a las definidas por CRISP-DM, pueden ser revisadas a mayor detalle por el lector en el ANEXO A. Traducción del modelo de referencia CRISP DM 1.0, para lo cual en cada caso se referencia la tarea específica.

Tabla 6. Actividades del proceso Iteración sobre CRISP-DM

Etapa 1: Comprensión del Negocio		
Actividad	**Responsable**	**Descripción detallada de la actividad**
Determinar los objetivos del negocio	Analista de Minería de Datos	Equivalente a la Tarea 1.1 de CRISP-DM Entender desde una perspectiva de negocio y dentro de su contexto, lo que el cliente quiere lograr. Equilibrando los posibles objetivos y las restricciones. Salidas posibles: ● Antecedentes ● Objetivos de negocio ● Criterios de éxito de negocio
Identificar datos relacionados disponibles en DWH	DWH Experto	Una vez comprendida el problema en términos de negocio, el DWH Experto identificará los datos disponibles en el sistema de Data Warehouse que estén relacionados con el problema planteado.
Identificar datos relacionados disponibles en sistemas transaccionales	DBA - Experto en esquema de datos	Una vez comprendida el problema en términos de negocio, el DBA o Experto en esquema de datos, identificará los datos disponibles en los diversos sistemas transaccionales que estén relacionados con el problema planteado.
Evaluar la situación	Analista de Minería de Datos	Equivalente a la Tarea 1.2 de CRISP-DM Comprender con un mayor nivel de detalle el problema, los recursos disponibles, restricciones y otros factores que definirán el objetivo del análisis y el plan de proyecto. Salidas posibles: ● Inventario de recursos ● Requerimientos, presunciones y asunciones ● Riesgos y contingencias
Determinar los objetivos de DM	Analista de Minería de Datos	Equivalente a la Tarea 1.3 de CRISP-DM Describir en términos técnicos (Minería de Datos), los objetivos del proyecto. Salidas posibles: ● Objetivos de la Minería de Datos. ● Criterios de éxito de la Minería de Datos.
Elaborar plan de proyecto	Analista de Minería de Datos	Equivalente a la Tarea 1.4 de CRISP-DM Definir el plan adecuado para alcanzar los objetivos de la Minería de Datos, se debe especificar los pasos, incluyendo selección de herramientas y técnicas. En este paso se verifica y ajusta la gestión de configuración.

		Salidas: ● Plan de proyecto.
Etapa 2: Comprensión de los datos		
Actividad	**Responsable**	**Descripción detallada de la actividad**
Solicitar datos iniciales	Analista de Minería de Datos	Definir una estructura en la que se solicitarán los datos identificados dentro del inventario de recursos. Salidas: ● Esquemas de datos iniciales
Recolectar y describir los datos solicitados	DWH Experto	Equivalente a las Tareas 2.1 y 2.2 de CRISP-DM Adquirir los datos listados como recursos del proyecto y que reposan en el DWH, y examinar las propiedades superficiales de los datos (número de registros, campos de las tablas) Salidas posibles: ● Informe de recolección de datos iniciales ● Informe de descripción de datos
Verificar la calidad de los datos	DWH Experto	Equivalente a las Tareas 2.4 de CRISP-DM Evaluar si los datos obtenidos poseen calidad: los datos son completos, están presentes errores (y qué tan frecuentes), existen valores perdidos, etc. Salidas posibles: ● Informe de calidad de datos
Exportar datos solicitados	DWH Experto	Exportar los datos según la estructura solicitada (si es conveniente y factible). Salida. ● Data Set
Recolectar y describir los datos solicitados	DBA - Experto en esquema de datos	Equivalente a las Tareas 2.1 y 2.2 de CRISP-DM Adquirir los datos listados como recursos del proyecto y que reposan en los sistemas transaccionales, examinar las propiedades superficiales de los datos (número de registros, campos de las tablas) Salidas posibles: ● Informe de recolección de datos iniciales ● Informe de descripción de datos
Verificar la calidad de los datos	DBA - Experto en esquema de datos	Equivalente a las Tareas 2.4 de CRISP-DM Evaluar si los datos obtenidos poseen calidad: los datos son completos, están presentes errores (y qué tan frecuentes), existen valores perdidos, etc.

		Salidas posibles: ● Informe de calidad de datos
Exportar datos solicitados	DBA - Experto en esquema de datos	Exportar los datos según la estructura solicitada (si es conveniente y factible). Salida. ● Data Set
Crear Data Mart	DWH Experto	Crear un Data Mart para el Data Set preparado por el DBA - Experto en esquema de datos, para ello será valioso contar con un adecuado Informe de recolección de datos iniciales donde se detallen las fuentes de los datos. Nota: ● En caso de que los datos proporcionados por el DBA - Experto en esquema de datos sean de interés para su inclusión dentro del DWH. Normalmente esto será adecuado si el proyecto podría ser repetido o generará un modelo con retroalimentación.
Explorar los datos	Analista de Minería de Datos	Equivalente a las Tareas 2.3 de CRISP-DM Consiste en analizar las preguntas de Minería de Datos usado técnicas de consulta, visualización y reportes, (previamente los datos deberían estar cargados en alguna herramienta para comprensión de datos). Esta actividad es sumamente importante puesto que generará conocimiento para la etapa de transformación. Salida: ● Informe de exploración de datos
Verificar la calidad de los datos	Analista de Minería de Datos	Equivalente a las Tareas 2.4 de CRISP-DM Evaluar si los datos obtenidos poseen calidad: los datos son completos, están presentes errores (y qué tan frecuentes), existen valores perdidos, etc. Salidas posibles: ● Informe de calidad de datos
Etapa 3: Preparación de los datos		
Actividad	**Responsable**	**Descripción detallada de la actividad**
Seleccionar los datos de análisis	Analista de Minería de Datos	Equivalente a la Tarea 3.1 de CRISP-DM Decidir qué datos serán utilizados para el análisis. En base a su importancia, calidad y restricciones. Nota: ● La selección se refiere tanto a las columnas a usar como a las filas o registros de una tabla.

		Salidas posibles: ● Razonamiento para inclusión o exclusión
Preparar los datos seleccionados	DWH Experto / DBA - Experto en esquema de datos	Adquirir los datos en base a la selección realizada.
Limpiar los datos	DWH Experto / DBA - Experto en esquema de datos	Mejorar la calidad de los datos a un nivel que permita la ejecución de las técnicas seleccionadas. Notas: ● Puede involucrar reemplazo de datos erróneos. Salidas posibles: ● Informe de limpieza de los datos.
Construir los datos	Analista de Minería de Datos	Construir datos a partir de operaciones especiales, por ejemplo para la creación de atributos derivados, o creación de nuevos registros o transformación de valores para atributos existentes. Salidas posibles: ● Atributos derivados ● Registros generados
Integrar los datos	Analista de Minería de Datos	Equivalente a la Tarea 3.4 de CRISP-DM Corresponde a la realización de operaciones para generar datos combinados a partir de tablas u otros registros. Salidas combinadas: ● Datos combinados
Actualizar Data Mart en DWH	DWH Experto	En base a las operaciones realizadas anteriormente se evalúa si los datos ahora forman parte del DWH, para que en este caso las operaciones realizadas puedan implementarse también dentro del DWH.
Formatear datos	Analista de Minería de Datos	Transformar sintácticamente los datos (estos no cambian su significado), es una operación realizada en base a las directrices dadas por la herramienta empleada más adelante para en el modelado.
Etapa 4: Modelado		
Actividad	**Responsable**	**Descripción detallada de la actividad**
Definir la técnica de modelado	Analista de Minería de Datos	Equivalente a la Tarea 4.1 de CRISP-DM Seleccionar la técnica de modelado real a usar, se refiere al algoritmo mismo que se empleará. Salidas posibles: ● Técnica de modelado ● Presunciones de modelado
Generar prueba	Analista de	Equivalente a la Tarea 4.2 de CRISP-DM

de diseño	Minería de Datos	Antes de construir el modelo, es necesario construir un procedimiento o mecanismo para probar la calidad y validez del modelo. Salidas posibles: ● Prueba de diseño
Construir el modelo	Analista de Minería de Datos	Equivalente a la Tarea 4.3 de CRISP-DM Ejecutar la herramienta para modelado con el Data Set preparado. Salidas: ● Configuración de parámetros ● Modelos ● Descripción de modelos
Evaluar el modelo	Analista de Minería de Datos	Equivalente a la Tarea 4.4 de CRISP-DM Interpretación y evaluación de los modelos según el conocimiento del dominio, los criterios de éxito de Minería de Datos y la prueba de diseño. Salidas: ● Evaluación de modelos ● Parámetro de ajustes revisados
Etapa 5: Evaluación		
Actividad	**Responsable**	**Descripción detallada de la actividad**
Evaluar los resultados	Analista de Minería de Datos	Equivalente a la Tarea 5.1 de CRISP-DM Se evalúa el grado en que el modelo cumple con los objetivos de negocio y si en este contexto los modelos poseen alguna deficiencia. Nota: ● Esta actividad es equivalente a la Evaluar el aporte del modelo realizada por el SME, en este caso se ha separado en dos, con la idea de retroalimentar la interpretación del Analista de Minería de Datos, con la interpretación de un experto.
Evaluar el aporte del modelo	SME - Experto en el tema	Equivalente a la Tarea 5.1 de CRISP-DM Se evalúa el grado en que el modelo cumple con los objetivos de negocio y si en este contexto los modelos poseen alguna deficiencia. Salidas posibles: ● Evaluación de los resultados de DM en relación a los criterios de éxito del negocio. ● Modelos aprobados

Revisar el proceso	Analista de Minería de Datos	Equivalente a la Tarea 5.2 de CRISP-DM Revisar en busca de omisiones en el proceso de Minería de Datos, en busca de asegurar la calidad. Salidas posibles: ● Revisión del proceso

ELABORACIÓN: Gustavo Cordero

5 APLICACIÓN DEL PROCESO EN TRES CASOS DE USO.

En el presente capítulo se procede a ejecutar el proceso metodológico definido en el capítulo anterior sobre tres problemas específicos de análisis de datos, estos son:

1. Identificar qué criterios entre asignaturas, calificaciones y datos de la ficha socioeconómica son los más relacionados al éxito académico en alumnos nuevos y analizar si estos son los mismos dependiendo de la facultades.
2. Cuán factible sería la construcción de una herramienta automática que permita anticipar casos de deserción sobre estudiantes que ingresan a la universidad.
3. Determinar si existe una relación entre el proceso de evaluación docente y la aprobación del alumno a la respectiva materia dictada por el docente evaluado.

En el caso del presente estudio, al corresponder a un proyecto de titulación cuyos objetivos estuvieron previamente definidos, se va a partir directamente desde las actividades propias del subproceso ***Gestión macro de proyectos de DM*** ya que las actividades de Levantar Requerimiento y Definir el equipo fueron previamente realizadas. El equipo definido para la ejecución de estos tres proyectos específicos consta de los siguientes participantes:

- Analista de Minería de Datos: Tesista.
- DWH Experto: técnico encargado de la construcción del sistema de Data Warehouse de la Universidad.
- DBA - Experto en esquemas de datos: especialista miembro de la DTIC

La especificación de la ejecución del proceso sobre los problemas planteados se hará a través de formularios a manera de tablas una por cada actividad del proyecto, como se puede observar en el ejemplo presentado en la Tabla 7. Si las actividades fueron realizadas en más de una ocasión, esto quiere decir en más de una iteración, esto será mencionado dentro de su descripción, y se entenderá que cada formulario corresponde a la última versión de la actividad especificada.

Tabla 7. Ejemplo de especificación de actividades seguidas en el proceso.

Proceso para la gestión de proyectos de Minería de Datos	
Actividad:	**Nombre de la actividad**
Descripción de la actividad realizada… Si la tarea fue realizada en más de una iteración se indicará brevemente este particular.	
Artefacto elaborado	Nombre del artefacto
Especificación del Artefacto...	

ELABORACIÓN: Gustavo Cordero

5.1 Problema 1: determinar los criterios más relacionados al éxito académico.

Definición del requerimiento: Identificar qué criterios entre asignaturas, calificaciones y datos de la ficha socioeconómica son los más relacionados al éxito académico en alumnos nuevos y analizar si estos son los mismos dependiendo de las facultades.

Nombre del proyecto: Selección de atributos relacionados al éxito académico.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico.
Actividad:	**Identificación del tipo de proyecto**
El problema definido busca que a partir de un conjunto de atributos disponibles se determine cuáles de ellos se encuentran más relacionados al éxito académico, por tanto el proyecto se refiere a un ejercicio de selección de atributos dentro del contexto de Minería de Datos.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Configurar el proyecto**
Una vez definido que el proyecto de negocio corresponde a un proyecto de Minería de Datos de selección de atributos, se procede a definir las actividades a ejecutar, para ello siguiendo la propuesta del modelo metodológico primero se planificará la ejecución de una iteración de experimentación.	

Nota: el artefacto producto de esta actividad fue modificado dado que posteriormente se encontró necesario que compartan el mismo data set que el problema siguiente, por tanto varios artefactos del proyecto 2 aplicarán al presente proyecto, por tanto sólo se procede a listar pero ya no serán incluidos nuevamente.	
Artefactos a emplear en Iteración sobre CRISP-DM (Experimental)	Artefactos a emplear en Iteración sobre CRISP-DM (Detallada)
1. Objetivos de Negocio 2. Inventarios de Recursos (Proyecto 2) 3. Objetivos de Minería de Datos y criterios de éxito 4. Esquema de datos iniciales (Proyecto 2) 5. Plan de inclusión o exclusión de datos (Proyecto 2) 6. Informe de resultados obtenidos por el modelo	1. Objetivos de Negocio 2. Inventarios de Recursos (Proyecto 2) 3. Objetivos de Minería de Datos y criterios de éxito (Proyecto 2) 4. Plan de proyecto 5. Esquema de datos iniciales (Proyecto 2) 6. Informe de exploración de datos (Proyecto 2) 7. Informe de calidad de los datos(Proyecto 2) 8. Plan de inclusión o exclusión de datos (Proyecto 2) 9. Informe de resultados obtenidos por el modelo

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Determinar los Objetivos de Negocio**
Partiendo de la definición del requerimiento lo que se busca son dos objetivos: 1. Determinar de entre los datos disponibles de los alumnos que ingresan a la universidad, cuáles de ellos se encuentran más relacionados al éxito académico. 2. Analizar los atributos seleccionados son los mismos para todas facultades o existe alguna diferenciación.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Identificar los datos relacionados y disponibles en el DHW**
Dado que el sistema de DWH se encuentra en fase de implementación, actualmente no existen datos disponibles y que sean de utilidad a esta problemática.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Identificar los datos relacionados y disponibles en sistemas transaccionales**
Al igual que el Problema 2: *Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad*, el presente problema está relacionado al éxito académico en estudiantes que ingresan a la	

universidad, por ello se consideró adecuado emplear el mismo data set para ambos problemas, y profundizar en el sobre los análisis por facultad y carrera como indica el problema planteado. Esto nos permitirá aportar en la selección de atributos para el problema 2.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Evaluar la situación**
Dado que el problema se refiere a alumnos nuevos la información disponible únicamente corresponde a la ficha socioeconómica que los alumnos deben llenar dentro del proceso de matriculación, y los datos de su matrícula como carrera y facultad. A continuación se listan el inventario de recursos.	
Inventario de recursos	• **Datos:** datos de la ficha socioeconómica de los alumnos, datos académicos de los alumnos. • **Personal:** como SME de negocio y Experto de Minería de Datos se cuenta del Director de la DTIC. Como expertos en los datos se cuenta con especialistas de la DTIC. • **Herramienta:** como herramienta a emplear se utilizarán Datacleaner, Spoon de pentaho, Weka.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Determinar los Objetivos de Minería de Datos**
En base a los requerimientos de negocio, los objetivos de Minería de Datos son los siguientes: • Aplicar una técnica de Minería de Datos para la selección de atributos que se encuentren más relacionados con el éxito académico. El **criterio de éxito** será obtener al menos un ranking que muestre los atributos que más se relacionan a la variable de interés (éxito académico).	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Elaborar Plan de Proyecto**
En esta actividad se procedió completar el artefacto Plan de Proyecto, el mismo dio inicio con la actividad de Configurar el Proyecto, en este punto se incluye la planificación de la duración de las etapas e iteraciones del proyecto.	

Artefacto	**Plan de Proyecto**

Gestión de configuración: hace referencia al artefacto generado en la actividad anterior
Configurar el proyecto.

Número de Iteraciones planificado: 1

- Experimental: iteración para realizar el análisis, no requiere de documentación considerable puesto que consiste más en un pequeño experimento de selección de atributos.

Fases a ejecutar: dado que el presente proyecto emplea los mismos datos que el proyecto 2, las fases de comprensión y preparación de los datos serán ejecutadas una sola vez para ambos proyectos, siendo estas registradas en el proyecto 2.

- 1. Comprensión del negocio (duración 1 días)
  - Entrada: requerimiento de negocio
  - Salida: objetivos de Minería de Datos
- 2. Modelamiento (duración 1 días)
  - Entrada: datos del Proyecto 1.
  - Salida: modelo de Minería de Datos.
- Evaluación (duración 1 días)
  - Entrada: salida de la fase anterior.
  - Salida: informe de los resultados obtenidos.

Duración total: 1 semana.

Riesgos identificados y plan de contingencia:

- Por el tamaño del proyecto los riegos no son representativos.

Herramienta a emplear: WEKA

Nota: el artefacto Plan de Proyecto si bien se inició con su elaboración en la primera iteración, con la especificación de la Configuración del Proyecto, posteriormente fue elaborado en la segunda.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Solicitar datos iniciales**

En esta actividad se estableció una matriz de ejemplo, para que en base a ella se generase un data set con los datos iniciales.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Definir la técnica de modelado**

En el caso del presente proyecto no se requiere la construcción de un modelo, en lugar de ello se requiere emplear una técnica de selección de atributos para lo cual se emplea algunos algoritmos evaluadores implementados dentro de la herramienta WEKA para comparar sus resultados todos ellos emplearán como método de búsqueda Ranker

- CorrelationAttributeEval
- InfoGainAttributeEval
- GainRatioAttributeEval

- OneRAttributeEval
- ReliefFAttributeEval

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Generar la prueba de diseño**

Para comprobar la validez de la selección de atributos se seleccionará al algoritmo con la mayor cantidad coincidentes en los primeros puestos con los otros rankings.

Una vez seleccionado el algoritmo se procederá a ejecutar nuevamente la selección de atributos en las Facultades y Carreras con mayor número de registros para analizar si los atributos más relacionados al éxito académico son diferentes en cada caso. Las facultades a analizar serán las siguientes:

1. FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS (# Instancias: 2018)
2. FACULTAD DE FILOSOFÍA, LETRAS Y CIENCIAS DE LA EDUCACIÓN (# Instancias: 1883)
3. FACULTAD DE CIENCIAS MÉDICAS (# Instancias: 1669)
4. FACULTAD DE INGENIERÍA (# Instancias: 984)

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Construir y evaluar el modelo**

Selección de atributos

A continuación son presentados los resultados de la ejecución de los algoritmos de selección de atributos:

CorrelationAttributeEval

Ranked attributes:

0.14748	**1 APORTES_IESS**
0.1473	**29 SEXO**
0.14138	**9 FICHA**
0.13936	33 TOTAL_EGRESOS
0.13926	**30 SUELDOS**
0.13557	**34 TOTAL_INGRESOS**
0.10716	**25 PERIODO_INICIAL**
0.09802	17 NRO_VEHICULOS
0.09361	**8 FACULTAD**
0.08863	35 ZONA_VIVIENDA
0.08388	7 ETNIA
0.07299	27 PROVINCIA_VIVIENDA_FAMILIAR
0.07173	10 IMPUESTO_PREDIAL
0.06986	12 MATERIALES_VIVIENDA
0.06886	2 AVALUO_ACUMULADO_VEHICULOS
0.06746	20 NUM_INTEGRANTES
0.06317	4 CARRERA
0.06002	21 NUM_LINEAS_TELEFONICAS
0.05695	3 CANTON_VIVIENDA_FAMILIAR
0.05203	11 IMPUESTO_RENTA
0.05141	14 NRO_PROPIEDADES_NO_RENTERAS
0.035	18 NUM_ESTUDIANTES_FAMILIA
0.03047	19 NUM_HIJOS_MEN6_ESTUDIANTE
0.02159	31 TENENCIA_VIVIENDA

InfoGainAttributeEval

Ranked attributes:

0.169203916	**4 CARRERA**
0.109477003	**25 PERIODO_INICIAL**
0.086928155	**8 FACULTAD**
0.030037446	**30 SUELDOS**
0.027124776	**3 CANTON_VIVIENDA_FAMILIAR**
0.026561301	**7 ETNIA**
0.022257231	**34 TOTAL_INGRESOS**
0.021903543	**9 FICHA**
0.019159749	**1 APORTES_IESS**
0.018565962	**27 PROVINCIA_VIVIENDA_FAMILIAR**
0.017093686	33 TOTAL_EGRESOS
0.015885009	29 SEXO
0.006921948	2 AVALUO_ACUMULADO_VEHICULOS
0.006875822	17 NRO_VEHICULOS
0.005910869	20 NUM_INTEGRANTES
0.005606558	10 IMPUESTO_PREDIAL
0.004593177	12 MATERIALES_VIVIENDA
0.003848073	35 ZONA_VIVIENDA
0.003371882	21 NUM_LINEAS_TELEFONICAS
0.003042136	31 TENENCIA_VIVIENDA
0.002268579	11 IMPUESTO_RENTA
0.002046476	13 MENSUAL_PAGO_ARRIENDO
0.001981482	18 NUM_ESTUDIANTES_FAMILIA
0.001851755	14 NRO_PROPIEDADES_NO_RENTERAS

0.01968 23 PAIS_VIVIENDA_FAMILIAR
0.01962 15 NRO_PROPIEDADES_RENTERAS
0.0178 13 MENSUAL_PAGO_ARRIENDO
0.01637 6 ESTUDIO_OTRA_CARRERA
0.0122 22 OTROS_INGRESOS
0.01091 16 NRO_PROPIEDADES_VACACIONALES
0.0081 28 RENTAS
0.00324 24 PENSIONES
0.003 26 PORCENTAJE_DISCAPACIDAD
0.00152 32 TIENECARNET

0.0007233 19 NUM_HIJOS_MEN6_ESTUDIANTE
0.000477533 23 PAIS_VIVIENDA_FAMILIAR
0.00038843 6 ESTUDIO_OTRA_CARRERA
0.000000299 32 TIENECARNET
0 24 PENSIONES
0 15 NRO_PROPIEDADES_RENTERAS
0 28 RENTAS
0 26 PORCENTAJE_DISCAPACIDAD
0 16 NRO_PROPIEDADES_VACACIONALES
0 22 OTROS_INGRESOS

Este algoritmo generó un ranking representativo con relación a los rankings de los otros algoritmos. (Sus 10 primeros atributos estuvieron también en los 10 primeros de al menos dos rankings)

GainRatioAttributeEva

Ranked attributes:
0.084151407 9 FICHA
0.039474774 23 PAIS_VIVIENDA_FAMILIAR
0.038189231 25 PERIODO_INICIAL
0.030863335 4 CARRERA
0.026525724 8 FACULTAD
0.022810118 27 PROVINCIA_VIVIENDA_FAMILIAR
0.020809079 7 ETNIA
0.01850509 3 CANTON_VIVIENDA_FAMILIAR
0.016014552 29 SEXO
0.015565855 33 TOTAL_EGRESOS
0.013674517 30 SUELDOS
0.013345877 1 APORTES_IESS
0.010108253 34 TOTAL_INGRESOS
0.009503326 2 AVALUO_ACUMULADO_VEHICULOS
0.007978362 17 NRO_VEHICULOS
0.007792705 20 NUM_INTEGRANTES
0.007361584 10 IMPUESTO_PREDIAL
0.005961524 35 ZONA_VIVIENDA
0.005843108 12 MATERIALES_VIVIENDA
0.004984711 11 IMPUESTO_RENTA
0.00363637 14 NRO_PROPIEDADES_NO_RENTERAS
0.003467886 21 NUM_LINEAS_TELEFONICAS
0.002489319 18 NUM_ESTUDIANTES_FAMILIA
0.002192051 13 MENSUAL_PAGO_ARRIENDO
0.001711137 31 TENENCIA_VIVIENDA
0.001294179 6 ESTUDIO_OTRA_CARRERA
0.001148421 19 NUM_HIJOS_MEN6_ESTUDIANTE
0.000000399 32 TIENECARNET
0 24 PENSIONES
0 15 NRO_PROPIEDADES_RENTERAS
0 28 RENTAS
0 26 PORCENTAJE_DISCAPACIDAD
0 16 NRO_PROPIEDADES_VACACIONALES
0 22 OTROS_INGRESOS

OneRAttributeEval

Ranked attributes:
74.3669 4 CARRERA
72.5661 8 FACULTAD
69.6023 3 CANTON_VIVIENDA_FAMILIAR
69.5648 27 PROVINCIA_VIVIENDA_FAMILIAR
69.5179 7 ETNIA
68.9458 30 SUELDOS
68.777 34 TOTAL_INGRESOS
68.6738 17 NRO_VEHICULOS
68.6081 22 OTROS_INGRESOS
68.5425 9 FICHA
68.5425 6 ESTUDIO_OTRA_CARRERA
68.5425 16 NRO_PROPIEDADES_VACACIONALES
68.5425 15 NRO_PROPIEDADES_RENTERAS
68.5425 14 NRO_PROPIEDADES_NO_RENTERAS
68.5425 12 MATERIALES_VIVIENDA
68.5425 35 ZONA_VIVIENDA
68.5425 18 NUM_ESTUDIANTES_FAMILIA
68.5425 19 NUM_HIJOS_MEN6_ESTUDIANTE
68.5425 31 TENENCIA_VIVIENDA
68.5425 20 NUM_INTEGRANTES
68.5425 32 TIENECARNET
68.5425 25 PERIODO_INICIAL
68.5425 29 SEXO
68.5425 21 NUM_LINEAS_TELEFONICAS
68.5331 24 PENSIONES
68.5144 23 PAIS_VIVIENDA_FAMILIAR
68.505 26 PORCENTAJE_DISCAPACIDAD
68.4487 28 RENTAS
68.4206 13 MENSUAL_PAGO_ARRIENDO
68.2893 1 APORTES_IESS
68.2048 10 IMPUESTO_PREDIAL
68.0173 11 IMPUESTO_RENTA
67.9985 2 AVALUO_ACUMULADO_VEHICULOS
67.7828 33 TOTAL_EGRESOS

ReliefFAttributeEval

Ranked attributes:
0.15795348 4 CARRERA
0.0993716 25 PERIODO_INICIAL
0.08745076 8 FACULTAD
0.06993997 7 ETNIA
0.03334271 9 FICHA
0.01614184 27 PROVINCIA_VIVIENDA_FAMILIAR
0.01187404 3 CANTON_VIVIENDA_FAMILIAR
0.01082349 29 SEXO

0.00438004 18 NUM_ESTUDIANTES_FAMILIA 0.00406265 22 OTROS_INGRESOS 0.00388191 20 NUM_INTEGRANTES 0.00359574 30 SUELDOS 0.00327331 19 NUM_HIJOS_MEN6_ESTUDIANTE 0.0031434 34 TOTAL_INGRESOS 0.0028798 1 APORTES_IESS 0.00202589 12 MATERIALES_VIVIENDA 0.00150311 33 TOTAL_EGRESOS 0.00077847 6 ESTUDIO_OTRA_CARRERA 0.00071412 24 PENSIONES 0.00057122 10 IMPUESTO_PREDIAL 0.00021572 17 NRO_VEHICULOS 0.00012193 35 ZONA_VIVIENDA 0.00011031 2 AVALUO_ACUMULADO_VEHICULOS 0.00002814 23 PAIS_VIVIENDA_FAMILIAR 0.00000804 15 NRO_PROPIEDADES_RENTERAS -0.00007972 28 RENTAS -0.00019037 11 IMPUESTO_RENTA -0.00020898 26 PORCENTAJE_DISCAPACIDAD -0.00023448 32 TIENECARNET -0.00026261 16 NRO_PROPIEDADES_VACACIONALES -0.00030482 21 NUM_LINEAS_TELEFONICAS -0.00112528 13 MENSUAL_PAGO_ARRIENDO -0.00226036 31 TENENCIA_VIVIENDA -0.00404239 14 NRO_PROPIEDADES_NO_RENTERAS	

Posterior a la actividad de ***Evaluar los Resultados***, fue necesario regresar a esta actividad para aplicar el algoritmo seleccionado (*InfoGainAttributeEval*), pero en este caso ejecutándose por cada facultad seleccionada y analizar si los atributos seleccionados presentan una variación, comparados con los atributos resultantes entre los diez primeros con el análisis realizado a todas las facultades, a continuación se detallan:

1. **4 CARRERA**
2. **25 PERIODO_INICIAL**
3. **8 FACULTAD**
4. **30 SUELDOS**
5. **3 CANTON_VIVIENDA_FAMILIAR**
6. **7 ETNIA**
7. **34 TOTAL_INGRESOS**
8. **9 FICHA**
9. **1 APORTES_IESS**
10. **27 PROVINCIA_VIVIENDA_FAMILIAR**

FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS	FACULTAD DE FILOSOFÍA, LETRAS Y CIENCIAS DE LA EDUCACIÓN
Ranked attributes: **0.0632578 4 CARREA** **0.0630556 25 PERIODO_INICIAL** **0.0240373 30 SUELDOS** **0.018222 34 TOTAL_INGRESOS** **0.0157957 3 CANTON_VIVIENDA_FAMILIAR** **0.0122267 29 SEXO** **0.0112244 9 FICHA** **0.0073275 31 TENENCIA_VIVIENDA** **0.0066593 12 MATERIALES_VIVIENDA** **0.006182 19 NUM_HIJOS_MEN6_ESTUDIANTE** 0.0057204 18 NUM_ESTUDIANTES_FAMILIA **0.0043129 7 ETNIA** **0.0034663 27 PROVINCIA_VIVIENDA_FAMILIAR** 0.003336 14 NRO_PROPIEDADES_NO_RENTERAS 0.000477 35 ZONA_VIVIENDA 0.0004735 6 ESTUDIO_OTRA_CARRERA 0.0000198 32 TIENECARNET	Ranked attributes: **0.266867402 4 CARRERA** **0.141959194 3 CANTON_VIVIENDA_FAMILIAR** **0.130888993 27 PROVINCIA_VIVIENDA_FAMILIAR** **0.118530711 7 ETNIA** **0.117576795 25 PERIODO_INICIAL** **0.06424072 12 MATERIALES_VIVIENDA** **0.03931719 35 ZONA_VIVIENDA** **0.037147323 34 TOTAL_INGRESOS** **0.036770175 9 FICHA** **0.022048004 33 TOTAL_EGRESOS** **0.017544736 30 SUELDOS** 0.01371088 20 NUM_INTEGRANTES 0.010691679 31 TENENCIA_VIVIENDA **0.009057723 1 APORTES_IESS** 0.008614474 13 MENSUAL_PAGO_ARRIENDO 0.00795191 18 NUM_ESTUDIANTES_FAMILIA 0.006791163 19 NUM_HIJOS_MEN6_ESTUDIANTE

0 8 FACULTAD 0 28 RENTAS 0 20 NUM_INTEGRANTES 0 10 IMPUESTO_PREDIAL 0 33 TOTAL_EGRESOS 0 2 AVALUO_ACUMULADO_VEHICULOS 0 26 PORCENTAJE_DISCAPACIDAD 0 11 IMPUESTO_RENTA 0 21 NUM_LINEAS_TELEFONICAS 0 16 NRO_PROPIEDADES_VACACIONALES 0 22 OTROS_INGRESOS 0 17 NRO_VEHICULOS 0 15 NRO_PROPIEDADES_RENTERAS 0 24 PENSIONES 0 23 PAIS_VIVIENDA_FAMILIAR 0 13 MENSUAL_PAGO_ARRIENDO **0 1 APORTES_IESS**	0.00386554 29 SEXO 0.000293857 23 PAIS_VIVIENDA_FAMILIAR 0.000023649 6 ESTUDIO_OTRA_CARRERA 0.000000629 32 TIENECARNET 0 26 PORCENTAJE_DISCAPACIDAD 0 2 AVALUO_ACUMULADO_VEHICULOS 0 28 RENTAS **0 8 FACULTAD** 0 21 NUM_LINEAS_TELEFONICAS 0 10 IMPUESTO_PREDIAL 0 22 OTROS_INGRESOS 0 17 NRO_VEHICULOS 0 16 NRO_PROPIEDADES_VACACIONALES 0 15 NRO_PROPIEDADES_RENTERAS 0 14 NRO_PROPIEDADES_NO_RENTERAS 0 24 PENSIONES 0 11 IMPUESTO_RENTA
FACULTAD DE CIENCIAS MÉDICAS Ranked attributes: **0.1544985 25 PERIODO_INICIAL** **0.06240802 34 TOTAL_INGRESOS** **0.05919287 30 SUELDOS** **0.04580344 1 APORTES_IESS** **0.04033535 9 FICHA** **0.03799323 33 TOTAL_EGRESOS** **0.03339996 7 ETNIA** **0.02487945 4 CARRERA** **0.02403288 3 CANTON_VIVIENDA_FAMILIAR** **0.01577777 20 NUM_INTEGRANTES** 0.00920684 11 IMPUESTO_RENTA **0.00864873 27 PROVINCIA_VIVIENDA_FAMILIAR** 0.00840821 17 NRO_VEHICULOS 0.0076285 21 NUM_LINEAS_TELEFONICAS 0.00606747 14 NRO_PROPIEDADES_NO_RENTERAS 0.00430141 29 SEXO 0.00327954 6 ESTUDIO_OTRA_CARRERA 0.0032536 31 TENENCIA_VIVIENDA 0.00285231 35 ZONA_VIVIENDA 0.00143719 12 MATERIALES_VIVIENDA 0.00000566 32 TIENECARNET 0 28 RENTAS 0 26 PORCENTAJE_DISCAPACIDAD 0 16 NRO_PROPIEDADES_VACACIONALES **0 8 FACULTAD** 0 2 AVALUO_ACUMULADO_VEHICULOS 0 24 PENSIONES 0 23 PAIS_VIVIENDA_FAMILIAR 0 15 NRO_PROPIEDADES_RENTERAS 0 10 IMPUESTO_PREDIAL 0 19 NUM_HIJOS_MEN6_ESTUDIANTE 0 13 MENSUAL_PAGO_ARRIENDO 0 22 OTROS_INGRESOS 0 18 NUM_ESTUDIANTES_FAMILIA	FACULTAD DE INGENIERÍA Ranked attributes: **0.0706345 25 PERIODO_INICIAL** **0.0290322 4 CARRERA** **0.0275416 18 NUM_ESTUDIANTES_FAMILIA** **0.0216553 3 CANTON_VIVIENDA_FAMILIAR** **0.0185813 7 ETNIA** **0.0130313 22 OTROS_INGRESOS** **0.0120449 2 AVALUO_ACUMULADO_VEHICULOS** **0.011825 6 ESTUDIO_OTRA_CARRERA** **0.0116977 9 FICHA** **0.0112612 27 PROVINCIA_VIVIENDA_FAMILIAR** 0.0087995 17 NRO_VEHICULOS 0.0086977 31 TENENCIA_VIVIENDA 0.0064308 12 MATERIALES_VIVIENDA 0.0025028 35 ZONA_VIVIENDA 0.0003383 29 SEXO 0.0000805 32 TIENECARNET 0 26 PORCENTAJE_DISCAPACIDAD 0 28 RENTAS **0 8 FACULTAD** 0 21 NUM_LINEAS_TELEFONICAS 0 20 NUM_INTEGRANTES **0 30 SUELDOS** 0 33 TOTAL_EGRESOS 0 10 IMPUESTO_PREDIAL 0 11 IMPUESTO_RENTA 0 24 PENSIONES 0 23 PAIS_VIVIENDA_FAMILIAR 0 19 NUM_HIJOS_MEN6_ESTUDIANTE **0 34 TOTAL_INGRESOS** 0 16 NRO_PROPIEDADES_VACACIONALES 0 13 MENSUAL_PAGO_ARRIENDO 0 15 NRO_PROPIEDADES_RENTERAS 0 14 NRO_PROPIEDADES_NO_RENTERAS **0 1 APORTES_IESS**

En azul se presentan los atributos que forman parte del top ten del ranking a nivel de todas las facultades, y en naranja los atributos que en el análisis por cada facultad se insertan dentro de los diez primeros.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Evaluar los resultados**

Luego de la ejecución de los algoritmos para selección de atributos, se procedió a determinar cuáles son los resultados comunes, es decir qué atributos fueron los que encabezaron las listas de los rankings de cada algoritmo ejecutado. Como se puede observar en los resultados indicados en el formulario anterior ***Construir y evaluar el modelo***, donde se han resaltado en color azul aquellos atributos que están dentro del top ten en tres o más rankings, con esta estrategia se puede ver que los atributos de **InfoGainAttributeEval**, son los que comparten mayor coincidencia con el resto de algoritmos, por lo que se puede indicar que los 10 atributos más relacionados al éxito académico (Egresamiento), son:

1. 4 CARRERA
2. 25 PERIODO_INICIAL
3. 8 FACULTAD
4. 30 SUELDOS
5. 3 CANTON_VIVIENDA_FAMILIAR
6. 7 ETNIA
7. 34 TOTAL_INGRESOS
8. 9 FICHA
9. 1 APORTES_IESS
10. 27 PROVINCIA_VIVIENDA_FAMILIAR

Luego de la ejecución del segundo análisis, sobre si existirían variaciones en el listado de atributos seleccionados, cuando se enfoca en facultades individuales. Se puede observar como muestran las tablas de la actividad ***Construir y evaluar el modelo***, que la mayoría de atributos corresponden a la lista del análisis general, sin embargo si existe una variación sobre ciertos atributos los que se procede a indicar:

- El atributo Facultad deja de ser de aporte, esto es lógico puesto que al ser el criterio del filtrado en este experimento, todas sus instancias son iguales en cada caso.
- En el análisis sobre la FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS, dentro de los diez primeros atributos presentados aparecen los siguientes: SEXO, TENENCIA_VIVIENDA, MATERIALES_VIVIENDA, NUM_HIJOS_MEN6_ESTUDIANTE, atributos coherentes con el hecho de que esta facultad oferta carreras en jornada vespertina, lo que posibilita una presencia mayor de estudiantes que trabajan y poseen sus propias familias.
- En los análisis sobre LA FACULTAD DE FILOSOFÍA, LETRAS Y CIENCIAS DE LA EDUCACIÓN, LA FACULTAD DE CIENCIAS MÉDICAS Y LA FACULTAD DE INGENIERÍA en cambio se insertan atributos como: TOTAL_EGRESOS, MATERIALES_VIVIENDA, ZONA_VIVIENDA, NUM_INTEGRANTES, NUM_ESTUDIANTES_FAMILIA, OTROS_INGRESOS, AVALUO_ACUMULADO_VEHICULOS, ESTUDIO_OTRA_CARRERA factores que están más relacionados a la condición socioeconómica de los alumnos.
- En términos generales se puede observar que los principales atributos CARRERA, PERIODO_INICIAL, FACULTAD siguen siendo los de mayor relación al Egresamiento.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Selección de atributos relacionados al éxito académico
Actividad:	**Evaluar el aporte del modelo**
Los resultados de los análisis son válidos puesto que se logró cumplir con ambos objetivos propuestos en el proyecto.	

5.2 Problema 2: determinar la factibilidad de construir una herramienta de predicción de deserción en alumnos nuevos.

Descripción del requerimiento: se requiere determinar si es factible la construcción de una herramienta automática que permita anticipar casos de probables desertores sobre estudiantes que ingresan a la universidad.

Nombre del proyecto: Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Identificación del tipo de proyecto**
El problema definido busca determinar si sería factible la implementación de una herramienta informática que permita reconocer de forma anticipada si un estudiante que ingresa a la universidad tiene alta probabilidad de desertar. Por tanto este corresponde a un problema de **Clasificación** dentro del contexto de Minería de Datos, donde se busca anticipar una variable (variable a predecir) mediante la disposición de los valores de otras variables (variables predictoras).	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Configurar el proyecto**
Una vez definido que el proyecto de negocio corresponde a un proyecto de Minería de Datos de clasificación, se procede a definir las actividades a ejecutar, para ello siguiendo la propuesta del modelo metodológico primero se planificará la ejecución de una iteración de experimentación.	
Artefactos a emplear en Iteración sobre CRISP-DM (Experimental)	Artefactos a emplear en Iteración sobre CRISP-DM (Detallada)
7. Objetivos de Negocio 8. Inventarios de Recursos 9. Objetivos de Minería de Datos y criterios de éxito 10. Esquema de datos iniciales 11. Plan de inclusión o exclusión de datos 12. Informe de resultados obtenidos por el modelo	10. Objetivos de Negocio 11. Inventarios de Recursos 12. Objetivos de Minería de Datos y criterios de éxito 13. Plan de proyecto 14. Esquema de datos iniciales 15. Informe de exploración de datos 16. Informe de calidad de los datos 17. Plan de inclusión o exclusión de datos 18. Informe de resultados obtenidos por el modelo

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Determinar los Objetivos de Negocio**
Partiendo de la definición del requerimiento, esta es bastante clara en términos de negocio, el objetivo es determinar si es posible, con los datos disponibles en los sistemas de la Universidad, construir una herramienta que permita la temprana identificación de estudiantes que ingresan y que poseen una alta probabilidad de	

desertar; puesto que al ser el objetivo aplicarse sobre estudiantes que están ingresando a la Universidad, los datos utilizables de ellos corresponden a su ficha socioeconómica y los datos de la carrera a la que se han matriculado.
Por tanto los objetivos son dos:

1. Determinar si es factible construir la herramienta de predicción que utilice como entrada los datos de la ficha socioeconómica y de la carrera.
2. Determinar la precisión que esta herramienta alcanzaría.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Identificar los datos relacionados y disponibles en el DHW**
Dado que el sistema de DWH se encuentra en fase de implementación, actualmente no existen datos disponibles y que sean de utilidad a esta problemática.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Identificar los datos relacionados y disponibles en sistemas transaccionales**
Como se pudo identificar en la determinación de los objetivos de negocio, la información que estaría disponible corresponden a los datos recabados de las fichas socioeconómicas de los alumnos y los académicos que al estar iniciando sus estudios únicamente corresponde a la información que identifica su carrera. En esta actividad se analiza con el encargado de la administración de las bases de datos, para verificar que estos dos tipos de información pueden ser obtenidos de los sistemas de la universidad, obteniéndose una respuesta afirmativa.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Evaluar la situación**
Dado que el conjunto de datos objeto de análisis fue identificado en las actividades anteriores, en la presente actividad se procede a definir un inventario de recursos disponibles.	
Inventario de recursos	• **Datos:** datos de la ficha socioeconómica de los alumnos, datos académicos de los alumnos. • **Personal:** como SME de negocio y Experto de Minería de Datos se cuenta del Director de la DTIC. Como expertos en los datos se cuenta con especialistas de la DTIC.

	• **Herramienta:** como herramienta a emplear se utilizarán Datacleaner, Spoon de pentaho, Weka.
Requerimientos, presunciones y restricciones	**Requerimientos - restricciones:** • Los datos manejados en el presente estudio no son públicos y no se puede hacer uso de los mismos sin autorización expresa. **Presunciones:** • Los datos correspondientes a la ficha socioeconómica y académica son completos y consistentes. • Se presume que es factible la construcción de un modelo predictivo de deserción.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Determinar los Objetivos de Minería de Datos**
En base a los requerimientos de negocio, el objetivo de Minería de Datos es el siguiente: • Construir un modelo de clasificación para predecir la deserción de un alumno que ingresa a la universidad basado en la información socioeconómica del mismo y sus datos académicos. El **criterio de éxito** será obtener un modelo que tenga una precisión superior al 75% sobre el grupo de datos de prueba.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Elaborar Plan de Proyecto**
En esta actividad se procedió completar el artefacto Plan de Proyecto, el mismo dio inicio con la actividad de Configurar el Proyecto, en este punto se incluye la planificación de la duración de las etapas e iteraciones del proyecto.	
Artefacto	**Plan de Proyecto**
Gestión de configuración: hace referencia al artefacto generado en la actividad anterior Configurar el proyecto. **Número de Iteraciones planificado**: 2 • Experimental: iteración para determinar la factibilidad del modelo. • Desarrollo: iteración completa a través de las fases de crisp-dm **Fases a ejecutar**: no se requiere el desarrollo de una fase de implementación. • 1. Comprensión del negocio (duración 5 días) ○ Entrada: requerimiento de negocio ○ Salida: objetivos de Minería de Datos	

- 2. Comprensión de los datos (duración 5 días)
  - Entrada: salida de la fase anterior.
  - Salida: inventario de recursos, plan de proyecto.
- 3. Preparación de los datos (duración 10 días)
  - Entrada: salida de la fase anterior.
  - Salida: data set base para el análisis.
- 4. Modelamiento (duración 5 días)
  - Entrada: salida de la fase anterior.
  - Salida: modelo de Minería de Datos.
- Evaluación (duración 2 días)
  - Entrada: salida de la fase anterior.
  - Salida: informe de los resultados obtenidos.

Duración total: 1 mes.

Riesgos identificados y plan de contingencia:

- Datos básicos para el análisis, que estén incompletos o de mala calidad: si este fuera el caso, antes de dar inicio a la segunda iteración se informará del particular al director departamental para que se evalúe la conveniencia del proyecto.

Nota: el artefacto Plan de Proyecto si bien se inició con su elaboración en la primera iteración, con la especificación de la Configuración del Proyecto, posteriormente fue elaborado en la segunda.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Solicitar datos iniciales**

En esta actividad se estableció una matriz de ejemplo, para que en base a ella se generase un data set con los datos iniciales.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Explorar y describir los datos solicitados**

Dado que en el Data Warehouse no existieron datos que aporten al problema planteado, esta actividad fue únicamente realizada en base a los sistemas transaccionales de la Universidad, la extracción de los datos fue realizada mediante consultas SQL a las bases de datos, la exploración en cambio fue realizada revisando algunos reportes propios de los sistemas, únicamente con la finalidad de verificar la completitud de los datos y el número de registros generados.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Verificar la calidad de los datos**

El DBA, en esta actividad realizó un proceso de revisión del tipo de datos registrados en las columnas disponibles, en busca de asegurar una buena calidad de los datos, para ello se utilizaron técnicas generales de limpieza de datos y eliminación de valores nulos.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Exportar datos solicitados**
El DBA, procedió a generar los Data Set solicitados en base a los formatos e indicaciones dadas por el analista. Los datos fueron exportados en formato de Excel.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Explorar y describir los datos solicitados**

En esta actividad la exploración de los datos fue realizada con las herramientas: Spoon (Pentaho) y Datacleaner, ambas tienen funcionalidades que permiten análisis de datos, sin embargo DataCleaner permite un análisis estadístico más rápido sobre los datos.

Se inició con el análisis de los datos de entrada, data set denominado Alumnos Nuevos, el cual consta de 26.615 observaciones o instancias, cada una correspondiente a los datos de un estudiante que ingresa a una determinada carrera en un período específico, posee 40 columnas o variables, las cuales corresponden a datos que identifican al estudiante, la carrera que cursa, información socioeconómica (tomada de las fichas socioeconómicas), y datos académicos como Promedio, número Total de asignaturas, y la variable más importante, la que el modelo deberá predecir Egresamiento, esta variable indica si el alumno ha egresado o no de la Universidad.
Este data set contiene la información de todos los estudiantes en el momento en que recién ingresaron a la Universidad, incluyendo registros desde el período Septiembre 2009 - Febrero 2010 hasta el período actual Marzo 2017 - Agosto 2017, incluyendo por lo tanto a los alumnos que ya Egresaron, a los que Desertaron y a los que están Cursando su carrera, debiéndose excluir del data set a los últimos, puestos que aun sobre ellos no se puede asegurar si van a egresar o desertar. Al no existir un atributo dentro del data set que permita diferenciar el último grupo, fue necesario obtener un listado de los alumnos que se han matriculado en el último período lectivo formando un data set denominado Alumnos Activos. Con ello se puede ejecutar la operación siguiente:

Alumnos Nuevos - Alumnos Activos = DS Egresados + Desertores.

Una vez realizado el proceso anterior el número de instancias se redujo a 10.662, y al hacerlo se eliminaron en buena medida varios outliers que se observaban en el data set original.
Este nuevo data set DS Egresados + Desertores constituirá los datos base para la construcción de modelo predictivo.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca

Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Verificar la calidad de los datos (Analista)**
Como técnica para identificar la calidad de los datos se procedió a utilizar diagramas de cajas o bigotes para cada atributo, estos diagramas usan el rango intercuartil (IQR) para resaltar los datos anómalos (cualquier instancia por encima de su límite superior o debajo de su límite inferior), con el objeto de que los mismos sean excluidos del data set a modelar. Sin embargo al analizar el número de observaciones que siguiendo esta técnica se incluirían, se hace evidente la pérdida de una gran cantidad de observaciones (63.65% de las observaciones), lo que provocaría que el modelo predictivo resultante sea muy rígido, por ello se ha ajustado los límites en busca de que el número de observaciones a excluir no supere un 2% cada variable considerada. La ejecución de esta técnica se la hizo en la actividad *Limpiar los datos.*	

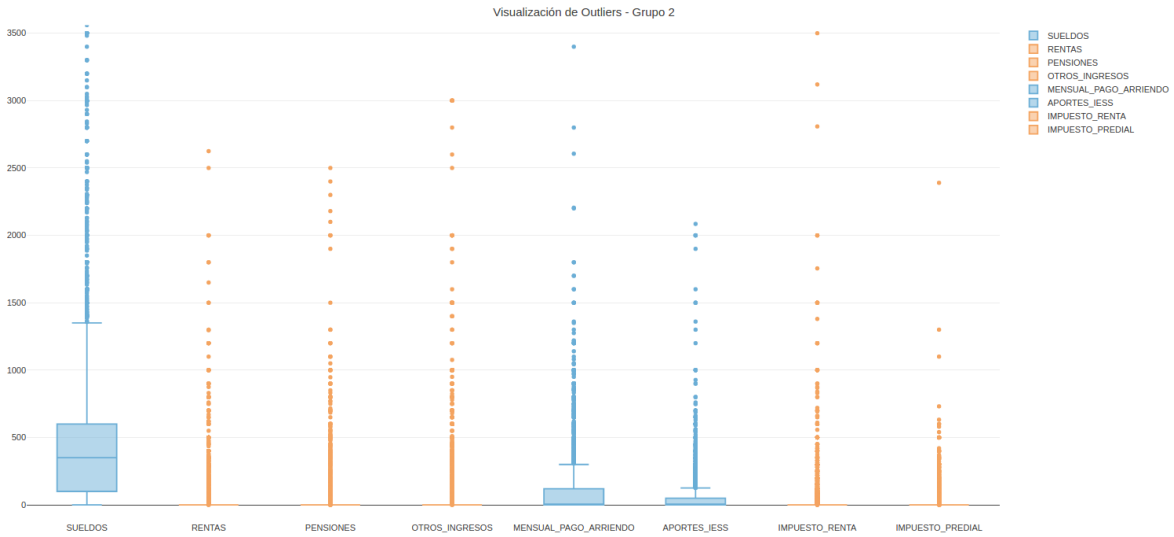
Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Seleccionar los datos de análisis**
Esta actividad se inició con una revisión detallada de cada campo en el data set DS Egresados + Desertores, para identificar si todas las variables deberían formar parte del data set a usarse en el modelamiento, en este punto se pudo anticipar que ciertas variables no aportarán en el modelamiento (clasificación). El primer caso corresponde a las columnas calculadas a partir de otras que también forman parte del data set. Otro caso es el de los campos relacionados a las calificaciones, estos datos que se obtienen tras la aprobación de al menos un nivel de una carrera, sin embargo al ser el objetivo del modelo predecir la deserción de los estudiantes que están ingresando, estas dos variables no deberían ser consideradas, puesto que los nuevos alumnos aún no poseen esta información. Por tanto las cuatro variables fueron eliminadas del data set.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Preparar datos seleccionados**
Siguiendo con la selección dada en la actividad anterior, en esta se procedió a eliminar las columnas no requeridas.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Limpiar los datos**
En esta actividad se ejecutaron las técnicas seleccionadas para depurar los datos y mejorar su calidad. Para ellos se construyó una transformación con la herramienta spoon para realizar la exclusión de outliers de una manera controlada. En este caso se logró eliminar los principales datos anómalos o outliers sin perder más allá	

de un 8.24% del total de observaciones, lo que es positivo puesto que permitirá que el modelo se entrene con mayor cantidad de observaciones.

Por otro lado los resultados del proyecto 1 permitieron validar el mecanismo empleado en este proyecto para la selección de atributos, puesto que las variables que se excluyeron mediante el análisis de outliers usando diagramas de caja y bigotes, fueron consistentes a las que estuvieron al último del ranking de selección de atributos. Como se muestra en la Figura 12.



Ranking de Selección de Atributos

Ranked attributes:
0.169203916 4 CARRERA
0.109477003 25 PERIODO_INICIAL
0.086928155 8 FACULTAD
0.030037446 30 SUELDOS
0.027124776 3 CANTON_VIVIENDA_FAMILIAR
0.026561301 7 ETNIA
0.021903543 9 FICHA
0.019159749 1 APORTES_IESS
0.018565962 27 PROVINCIA_VIVIENDA_FAMILIAR
0.015885009 29 SEXO
0.006921948 2 AVALUO_ACUMULADO_VEHICULOS
0.006875822 17 NRO_VEHICULOS
0.005910869 20 NUM_INTEGRANTES
0.005606558 10 IMPUESTO_PREDIAL
0.004593177 12 MATERIALES_VIVIENDA
0.003848073 35 ZONA_VIVIENDA
0.003371882 21 NUM_LINEAS_TELEFONICAS
0.003042136 31 TENENCIA_VIVIENDA
0.00255115 33 TIENE_TVCABLE
0.002268579 11 IMPUESTO_RENTA
0.002046476 13 MENSUAL_PAGO_ARRIENDO
0.001981482 18 NUM_ESTUDIANTES_FAMILIA
0.001851755 14 NRO_PROPIEDADES_NO_RENTERAS
0.0007233 19 NUM_HIJOS_MEN6_ESTUDIANTE
0.000477533 23 PAÍS VIVIENDA FAMILIAR
0.00038843 6 ESTUDIO_OTRA_CARRERA
0.000000323 34 TIPO DISCAPACIDAD
0.000000299 32 TIENECARNET
0 24 PENSIONES [X]
0 16 NRO_PROPIEDADES_VACACIONALES [X]
0 15 NRO_PROPIEDADES_RENTERAS [X]
0 26 PORCENTAJE_DISCAPACIDAD [X]
0 28 RENTAS [X]
0 22 OTROS_INGRESOS [X]

Figura 12. Diagramas de caja y bigotes vs Ranking de selección de atributos.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Construir e integrar los datos**

Dado que los datos fueron obtenidos mediante una misma fuente de consulta, no fue necesario realizar trabajos adicionales de integración y al ser la principal fuente de análisis los datos de la ficha socioeconómica, estos no requirieron de integración.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Formatear los datos**

En esta etapa se realizan varias pruebas usando la herramienta Weka, donde se utilizan distintas combinaciones de filtros de discretización y normalización. Con el objetivo de realizar las pruebas de clasificación también sobre un algoritmo de RNA (Redes Neuronales Artificiales), en esta etapa se elaboraron dos tipos de transformaciones: la primera que consistió en transformar las variables categóricas en números enteros del 1 en adelante (según el número de etiquetas que posee cada variable), y la segunda, discretizar las variables numéricas para luego igualmente transformar cada etiqueta resultante en números enteros.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Definir la técnica de modelado**

Para la construcción y validación del modelo se emplearán dos técnicas, por un lado SVM o Support Vector Machine y Perceptrón Multicapa (Redes Neuronales Artificiales), dada que ambas han sido utilizadas en problemas similares de aplicación de EDM (Minería de Datos Educacional).

Support Vector Machine: Es un modelo de aprendizaje supervisado, que dado un conjunto de ejemplos de entrenamiento (muestras) puede etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra, se aplican con frecuencia a problemas de clasificación y regresión (Patki et al., 2013). Con respecto a esta técnica el algoritmo empleado dentro de la herramienta de Minería de Datos WEKA corresponde a SMO empleando el kernel RBFKernel.

Redes Neuronales Artificiales RNA: este enfoque se basa en el funcionamiento de las redes neuronales biológicas, a términos generales las redes neuronales son conjuntos de unidades de entrada / salida conectadas, donde cada conexión entre dos neuronas posee un peso asociado, esta característica es la que permite que los métodos de clasificación basados en redes neuronales artificiales pueda ajustar los pesos de las interconexiones y con ello aprender de los datos de entrenamiento. El algoritmo que se utilizó corresponde al de Perceptrón Multicapa, este algoritmo corresponde a una red neuronal de múltiples capas completamente conectadas; esto significa que la salida de cada entrada y las neuronas ocultas es distribuida a todas las neuronas de la siguiente capa (Gupta, 2013). Toda red neuronal posee una capa de entrada y otra de salida, pero el número de capas ocultas puede variar. El algoritmo empleado dentro de la herramienta de Minería de Datos WEKA corresponde a MultilayerPerceptron (MLP).

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad

Actividad:	**Generar la prueba de diseño**

Para comprobar la validez del modelo se emplearán las tasas de instancias clasificadas correctamente o erróneamente.

Por otro lado para la elaboración del modelo basado en la técnica SVM (SMO) se utilizará la funcionalidad propia de WEKA para formar un grupo de entrenamiento con el 85% del data set y el 15% restante será corresponderá al grupo de prueba.

En cambio en la generación del modelo basado en RNA (MLP) variamos el tamaño del grupo de entrenamiento a 80% y 20% para el grupo de prueba.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Construir y evaluar el modelo**

Clasificación con SVN:

Resultado de la clasificación con SVM

Criterio	Resultado
Kernel Usado	K(x,y) = e^-(0.01* <x-y,x-y>^2)
Tiempo de construcción del modelo	250.51 segundos
Número de vectores de soporte	5.561
Número de instancias de prueba	1.599
Instancias clasificadas correctamente	1.306 **81.676 %**
Instancias clasificadas erróneamente	293 **18.324 %**
Error absoluto medio	0.1832
Raíz del error cuadrático medio	0.4281
Error relativo absoluto	42.4338%
Raíz del error cuadrático relativo absoluto	92.0382%

Matriz de confusión

a	b	<-- classified as
295	211 |	a = SI **(58.3%)**
82	1011 |	b = NO **(92.5%)**

Tras la ejecución del modelo construido se puede observar que a nivel general se obtiene un 82% de instancias correctamente clasificadas y un 18% de erróneas (usando el conjunto de pruebas). Revisando la matriz de confusión se puede detectar que el modelo es mucho más efectivo clasificando la deserción con un 92.5% de precisión, que el egresamiento con un 58.3%. ***Estos resultados implican una precisión satisfactoria del modelo para el objetivo propuesto, por tanto el modelo es apto para ser utilizado en la clasificación (predicción) de deserción en un grupo nuevo de alumnos.***

Nota:

Con este algoritmo se aplicaron distintas combinaciones de filtros de discretización y normalización, para medir el impacto de aplicarlos al data set de entrenamiento del modelo, sin embargo al comparar los resultados de la clasificación no se evidencia una mejora representativa al usar estas técnicas (mejora del 0.1877%).

Clasificación con RNA:

Resultado de Clasificación MLP entrenado con datos numéricos

Criterio		Resultado	Matriz de confusión		
Número de instancias	1599		a	b	<-- classified as
Instancias clasificadas correctamente	1263	**78.98 %**	335	171	a = SI **(66,2%)**
Instancias clasificadas erróneamente	336	**21.01 %**	165	928	b = NO **(85%)**
Error absoluto medio	0.2627				
Raíz del error cuadrático medio	0.3872				

Con la última técnica se obtuvo un 78.98% de instancias correctamente clasificadas y un 21.01% de erróneas. Revisando la matriz de confusión se puede detectar al igual que en modelo anterior, el modelo es mucho más efectivo clasificando la deserción con un 85% de precisión, que el egresamiento con un 66.2%, Comparado con el anterior este modelo mejoró su precisión de Egresamiento pero empeoró con relación a la deserción.

Nota:

Para probar la respuesta de esta técnica se hicieron pruebas generando el modelo sobre el data set original y en otra prueba se transformaron las variables previamente a numéricas, lo cual mejoró el tiempo de construcción del modelo pero las tasas fueron similares en ambos casos.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Evaluar los resultados**

Al evaluar los modelos generados con sus respectivos grupos de prueba, muestran una precisión general (Clasificar egresamiento o deserción) del 81.67% en el caso de SVM (SMO) y del 78.98% en el caso de MLP (RNA). Siendo el algoritmo de Support Vector Machine o SVM más eficiente que Perceptrón Multicapa en un 2.69%. Sin embargo, al ser el objetivo del presente problema la identificación de estudiantes con probabilidad de deserción, el análisis se puede enfocar en los resultado alcanzados puntualmente al predecir deserción (Egresamiento = No), en cuyo caso los resultados son mejores, en SVM se alcanza un 92.5% (1011/1093 fueron clasificados correctamente) y en MLP un 85% (928/1093 fueron clasificados correctamente). Se puede por tanto concluir que, el modelo construido posee una precisión satisfactoria, lo cual permitirá una identificación temprana de alumnos con riesgo de desertar. Se confirma también la adecuada aplicabilidad de estas técnicas a problemas de clasificación similares.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Evaluar el aporte del modelo**

Una vez obtenidos los resultados del modelo, estos son analizados en conjunto con el experto en el negocio, con quien se llega a determinar que el modelo aporta contundentemente para satisfacer la pregunta objeto del

presente proyecto, y de esta manera dar una respuesta afirmativa sobre **que es completamente factible la construcción de una herramienta (modelo) para predecir los casos de deserción estudiantil sobre alumnos que ingresan a la universidad.** Se recomienda sin embargo la ejecución de un nuevo proyecto en el que se incluyan más variables de análisis y que se incluya una fase de despliegue donde implementar el modelo dentro de un sistema operacional de la universidad que permita de manera automática determinar estos casos probables de deserción.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Revisar el proceso**
En la revisión al proceso se verifica la validez del estudio realizado, sin embargo se hace evidente en los datos analizados, que sería de mucho aporte en futuras instancias repetir el experimento con la inclusión de otro tipo de variables como: psicológicas, académicas (preuniversitarias), motivacionales, económicas (más actualizadas), sociales, etc.	

5.3 Problema 3: determinar si existe relación entre la evaluación docente y la aprobación estudiantil de una materia.

Descripción del requerimiento: Determinar si existe una relación entre el proceso de evaluación docente y la aprobación de los alumnos a la respectiva materia dictada por el docente.

Descripción del requerimiento: se requiere determinar si existe una relación entre las calificaciones obtenidas por un estudiante en una determinada materia y la evaluación realizada al docente que la impartió.

Nombre del proyecto: Análisis de grupos para identificar la relación entre el aprovechamiento estudiantil y la evaluación docente.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Identificación del tipo de proyecto**
El problema definido busca determinar qué tipo de relación existe entre el aprovechamiento del estudiante y la evaluación que el docente ha recibido por parte de sus alumnos, por tanto lo que se busca es determinar si en los datos existen agrupaciones que permitan inferir algún tipo de relación entre estas dos variables. Por lo tanto, la técnica de Minería de Datos a emplear corresponde a Clustering.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Configurar el proyecto**
Una vez definido que el objetivo de negocio corresponde a un proyecto de Minería de Datos de clustering, se procede a definir las actividades a ejecutar, para ello siguiendo la propuesta del modelo metodológico, primero se planificará la ejecución de una iteración de experimentación.	
Artefactos a emplear en Iteración sobre CRISP-DM (Experimental)	Artefactos a emplear en Iteración sobre CRISP-DM (Detallada)
13. Objetivos de Negocio 14. Inventarios de Recursos 15. Objetivos de Minería de Datos y criterios de éxito 16. Esquema de datos iniciales 17. Plan de inclusión o exclusión de datos 18. Informe de resultados obtenidos por el modelo	19. Objetivos de Negocio 20. Inventarios de Recursos 21. Objetivos de Minería de Datos y criterios de éxito 22. Plan de proyecto 23. Esquema de datos iniciales 24. Informe de exploración de datos 25. Informe de calidad de los datos 26. Plan de inclusión o exclusión de datos 27. Informe de resultados obtenidos por el modelo

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Determinar los Objetivos de Negocio**
Partiendo de la definición del requerimiento, esta es bastante clara en términos de negocio, el objetivo es determinar si existe relación entre la evaluación docente y el aprovechamiento estudiantil, identificando agrupaciones en los sujetos del estudio.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Identificar los datos relacionados y disponibles en el DHW**
Dado que el sistema de Data Warehouse se ha implementado cubos para el análisis de la evaluación docente, si existe cierta información disponible que se encuentra relacionada al problema planteado. Esta información corresponde a la calificación de un docente por función, período y cuestionario, como se muestra en la imagen siguiente.	

Docente: [redacted]
Facultad:
Escuela:

CUESTIONARIO	FUNCION		FUNCION DOCENCIA		TOTAL	TOTAL PONDERADO	
	VALOR	PORCENTAJE	VALOR	PORCENTAJE		VALOR	PORCENTAJE
AUTOEVALUACION DOCENTE - 2011	3,80 / 3,80	100,00%	6,28 / 6,40	98,12%	10,08 / 10,20	19,76 / 20,00	98,12%
CUESTIONARIO DE OPINION PARA ESTUDIANTES			124,38 / 149,76	83,05%	124,38 / 149,76	43,85 / 50,00	83,05%
CUESTIONARIO PARA AUTORIDADES ACADÉMICAS - 2011	5,72 / 5,72	100,00%	9,60 / 9,60	100,00%	15,32 / 15,32	30,00 / 30,00	100,00%
TOTAL					149,78 / 175,28	91,29 / 100,00	

Figura 13. Reporte de Evaluación de un Docente

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Identificar los datos relacionados y disponibles en sistemas transaccionales**
Como se pudo identificar en la determinación de los objetivos de negocio, la información que estaría disponible corresponden a los datos las notas de aprovechamiento de los estudiantes por un lado, y la evaluación realizada al docente por parte de los alumnos. En esta actividad se analiza con el encargado de la administración de las bases de datos, para verificar que estos dos tipos de información pueden ser obtenidos de los sistemas de la universidad, de esto se obtuvo una respuesta afirmativa.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Evaluar la situación**
Dado que el conjunto de datos objeto de análisis fue identificado en el las actividades anteriores, en la presente actividad se procede a definir el un inventario de recursos disponibles y	
Inventario de recursos	• **Datos:** aprovechamiento de los alumnos (notas por alumno y materia), evaluación a docentes: calificación por cuestionario, por función y por período de evaluación (1 evaluación cada dos períodos académicos), • **Personal:** como SME de negocio y Experto de Minería de Datos se cuenta del Director de la DTIC. Como expertos en los datos se cuenta con especialistas de la DTIC. • **Herramienta:** como herramienta a emplear se utilizarán Datacleaner, Spoon de pentaho, Weka, Rapidminer Studio Educational.
Requerimientos, presunciones y restricciones	**Requerimientos - restricciones:**

	• Los datos manejados en el presente estudio no son públicos y no se puede hacer uso de los mismos sin autorización expresa. **Presunciones:** • Se presume es factible llegar a obtener la evaluación que cada estudiante ha realizado sobre el docente de cada materia.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Determinar los Objetivos de Minería de Datos**
En base a los requerimientos de negocio, el objetivo de Minería de Datos es el siguiente: • Analizar la agrupación existente entre los datos de Aprovechamiento de los alumnos y la Evaluación Docente. El **criterio de éxito** será el contar al final de proyecto con una interpretación de la relación que se evidencie entre estas dos variables.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Elaborar Plan de Proyecto**
En esta actividad se procedió completar el artefacto Plan de Proyecto, el mismo dio inicio con la actividad de Configurar el Proyecto, en este punto se incluye la planificación de la duración de las etapas e iteraciones del proyecto.	
Artefacto	**Plan de Proyecto**

Gestión de configuración: hace referencia al artefacto generado en la actividad anterior Configurar el proyecto.
 	
Número de Iteraciones planificado: 2
• Experimental: iteración para determinar la factibilidad del modelo.
• Desarrollo: iteración completa a través de las fases de crisp-dm
Fases a ejecutar: no se requiere el desarrollo de una fase de implementación.
• 1. Comprensión del negocio (duración 5 días)
  ○ Entrada: requerimiento de negocio
  ○ Salida: objetivos de Minería de Datos
• 2. Comprensión de los datos (duración 10 días)
  ○ Entrada: salida de la fase anterior.
  ○ Salida: inventario de recursos, plan de proyecto.
• 3. Preparación de los datos (duración 10 días)
  ○ Entrada: salida de la fase anterior.
  ○ Salida: dataset base para el análisis.
• 4. Modelamiento (duración 5 días)
  ○ Entrada: salida de la fase anterior.	

- Salida: modelo de Minería de Datos.
- Evaluación (duración 2 días)
  - Entrada: salida de la fase anterior.
  - Salida: informe de los resultados obtenidos.

Duración total: 1 mes.

Riesgos identificados y plan de contingencia:

- Datos básicos para el análisis que estén incompletos o de mala calidad: si este fuera el caso, antes de dar inicio a la segunda iteración se informará del particular al director departamental para que se evalúe la conveniencia del proyecto.

Nota: el artefacto Plan de Proyecto si bien se inició con su elaboración en la primera iteración, con la especificación de la Configuración del Proyecto, posteriormente fue elaborado en la segunda.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Solicitar datos iniciales**

En esta actividad se estableció un diagrama entidad relación de ejemplo (Figura 14), para que en base a ella se generase un data set con los datos iniciales.

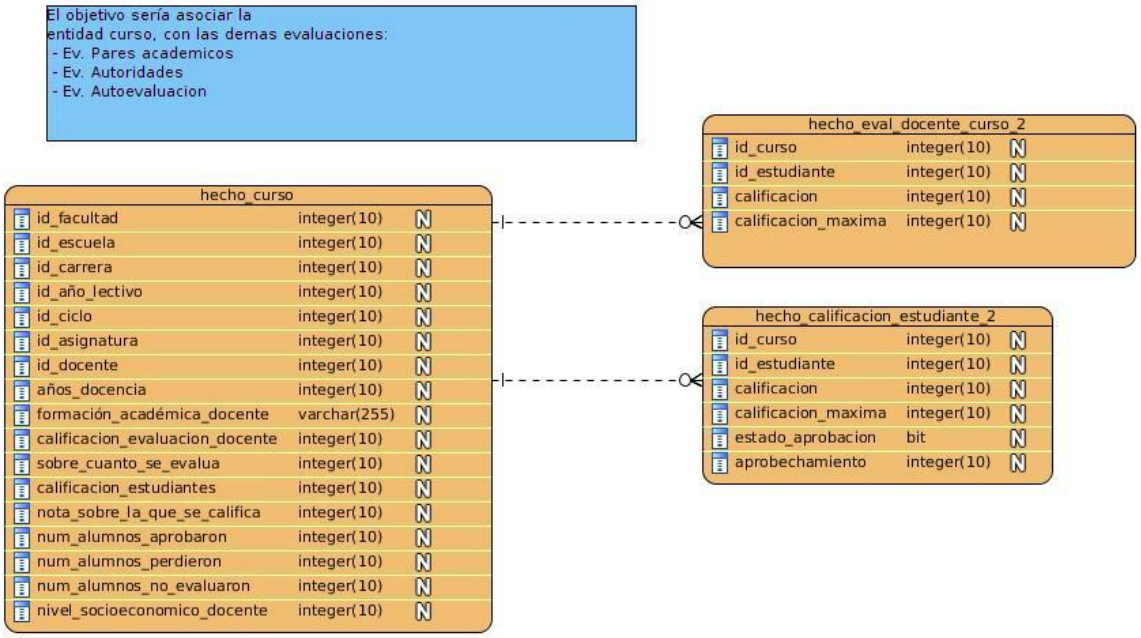


Figura 14. Diagrama entidad - relación de los datos requeridos.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente

Actividad:	**Explorar y describir los datos solicitados**
Los datos disponibles en el Data Warehouse (DWH) fueron explorados mediante revisión del esquema de la base de datos del DWH, y mediante consultas SQL, encontrándose que los datos correspondientes al aprovechamiento de los alumnos, era parcialmente consistente con la estructura requerida (Calificación por estudiante, carrera, materia, docente, período lectivo, y nivel o ciclo). Por otro lado, la información sobre Evaluación Docente no se encontró especificada con relación a cada estudiante, sino esta se estuvo definida como una calificación por Docente por período de evaluación y por función, lo cual no iba a permitir un adecuado análisis. Nota: En la ejecución del presente proyecto, los roles de **DBA** y **DWH - Experto** fueron ejecutados por el mismo técnico, por este motivo las actividades de Explorar y describir los datos solicitados, Verificar la calidad de los datos, Exportar datos solicitados, Preparar datos seleccionados y Limpiar los datos, no estarán diferenciadas por el rol del responsable es decir serán hechas una sola vez.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Verificar la calidad de los datos**
El DWH - Experto, en esta actividad realizó un proceso de revisión del tipo de datos registrados correspondientes a la información de aprovechamiento del estudiante, en busca de asegurar una buena calidad de los datos, para ello se utilizaron técnicas generales de limpieza de datos, eliminación de valores null y outliers.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Exportar datos solicitados**
Para poder obtener los datos sobre las evaluaciones a los docentes fue necesario desarrollar un subproyecto para la adquisición de estos datos puesto que estos debieron ser calculados a partir de las calificaciones de cada pregunta de los respectivos cuestionarios llenados por los estudiantes, realizándose varias iteraciones sobre este punto hasta finalmente construir un conjunto de tablas dentro del data warehouse con la información requerida.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Seleccionar los datos de análisis**
En actividad se inició con una revisión detallada de los campos disponibles en las tabla creadas, para identificar si todas las variables deberían formar parte del dataset a usarse en la clusterización, en una primera instancia se incluyeron las siguientes variables:	

- VECES_CURSA: número de veces que el alumno se ha matriculado en esa materia.
- ESTADO_APROBACION: estado de aprobación o reprobación de la materia.
- NOTA_APROVECHAMIENTO: calificación del estudiante.
- PROMEDIO_EVALUACION
- CARRERA_ID: identificador de la carrera del estudiante
- CARRERA_DESC: nombre de la carrera
- ASIGNA_ID: identificador de la materia
- NIVEL_NUMERO: ciclo en el que se toma la asignatura, depende de la carrera del estudiante.
- DOCENTE_ID: identificación del docente (dato anonimizado)
- PERLEC_ABREV: referencia del período lectivo
- estudiante_genero: género del estudiante
- docente_genero: género del docente

Posteriormente a las pruebas realizadas con los algoritmos de clustering, las variables que permitieron observar agrupaciones de mejor manera corresponden a:

- NOTA_APROVECHAMIENTO
- PROMEDIO_EVALUACION
- NIVEL_NUMERO

.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Preparar datos seleccionados**
Siguiendo con la selección dada en la actividad anterior, en esta se procedió a eliminar las columnas no requeridas.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Limpiar los datos**
En esta actividad se ejecutaron las técnicas seleccionadas para depurar los datos y mejorar su calidad. Para ello empleando la herramienta RapidMiner se procedió a identificar aquellas variables que presentaban datos perdidos para eliminarlos o reemplazarlos.	

Tabla 8. Estadísticas de los datos originales

ESTUDIANTE_ID	Integer	458	Min 221487	Max 95060309141	Average 287550399.530
NIVEL_NUMERO	Integer	0	Min -1	Max 12	Average 4.887
PK_CURSO	Integer	0	Min 4	Max 66002	Average 30251.130
VECES_CURSA	Integer	0	Min 1	Max 3	Average 1.069
ESTADO_APROBACION	Polynominal	0	Least RR (1)	Most AP (367748)	Values AP (367748), RE (37640), ...[5 more]
NOTA_APROVECHAMIENTO	Integer	6773	Min 0	Max 94375	Average 131.390
BAN_EVALUACION	Polynominal	0	Least N (116774)	Most S (293440)	Values S (293440), N (116774)
CUESTI_ID	Polynominal	0	Least 152 (23506)	Most 174 (135057)	Values 174 (135057), NULL (116774), ...[3 more]
NOTA_CUESTIONARIO	Polynominal	0	Least 99.93 (1)	Most NULL (116774)	Values NULL (116774), 49.92 (38592), ...[10781 more]
NOTA_CUESTIONARIO_TOTAL	Numeric	116774	Min 49.920	Max 440	Average 71.829
PROMEDIO_EVALUACION	Numeric	116774	Min 19.790	Max 100	Average 88.094

FUENTE: RapidMiner
ELABORACIÓN: Gustavo Cordero

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Construir e integrar los datos**
Para la integración de datos fue necesario obtener los datos de los sistemas académico y evaluación docente, estos fueron introducidos al sistema de Data Warehouse en una estructura de varias tablas (hechos y dimensiones). Para la aplicación de los algoritmos de Minería de Datos (Clustering) fue necesarios exportar los datos a una estructura de un único data set.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Formatear los datos**
Los datos preparados previamente fueron en su mayoría numéricos y sin la presencia de valores nulos, por lo que no fue necesaria la realización de operaciones adicionales para su formateo.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Definir la técnica de modelado**

En base a la revisión de los algoritmos disponibles por las herramientas de Minería de Datos se seleccionaron los algoritmos de: K-means (el número de agrupaciones k es definido como parámetro de inicio) y X-Means (Determina el número correcto de agrupaciones k, en base a heurísticas).

Se emplean las dos técnicas anteriormente indicadas por su simplicidad computacional, dado que se trató de emplear el algoritmo SVC (Support Vector Clustering) sin embargo no se obtuvo ningún resultado tras 15 horas de procesamiento.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Generar la prueba de diseño**
La prueba de la aplicación consistirá en la verificación de identificación de grupos similares entre las distintas técnicas y herramientas.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Construir y evaluar el modelo**

Al ejecutar el agrupamiento (clustering) con ambas técnicas se obtuvieron resultados muy similares, el mejor agrupamiento se realizó con 4 grupos (eso se confirmó al ejecutar X-Means y obtener que el número óptimo de clústeres fue 4), a continuación se presentan los resultados obtenidos.

Tabla 9. Centroides

Atributo	cluster_0	cluster_1	cluster_2	cluster_3
NIVEL_NUMERO	4.317	5.344	3.367	4.928
NOTA_APROVECHAMIENTO	65.344	85.743	34.717	73.476
PROMEDIO_EVALUACION	94.213	93.962	65.366	58.296

FUENTE: RapidMiner
ELABORACIÓN: Gustavo Cordero

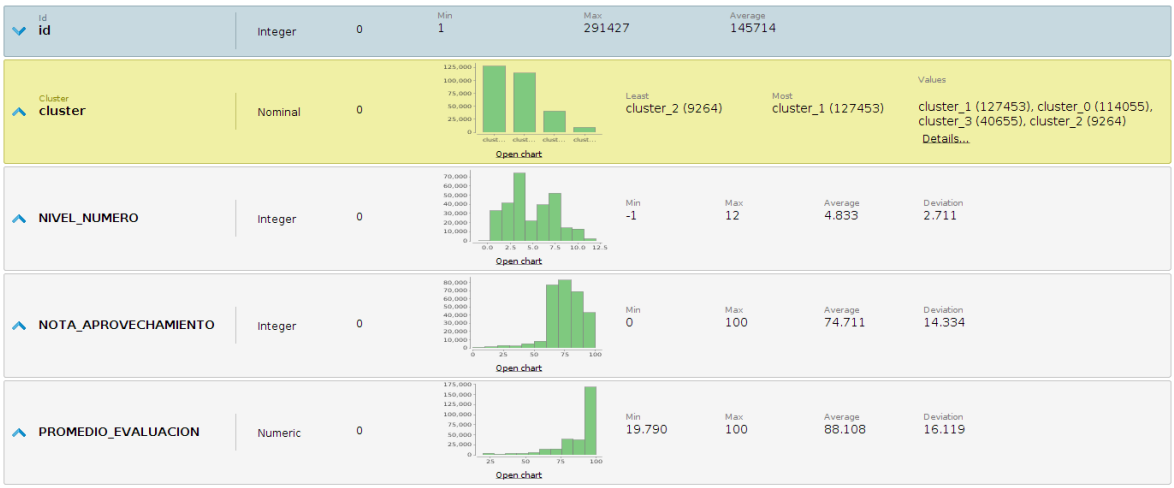


Figura 15. Estadísticas de los atributos finales del data set

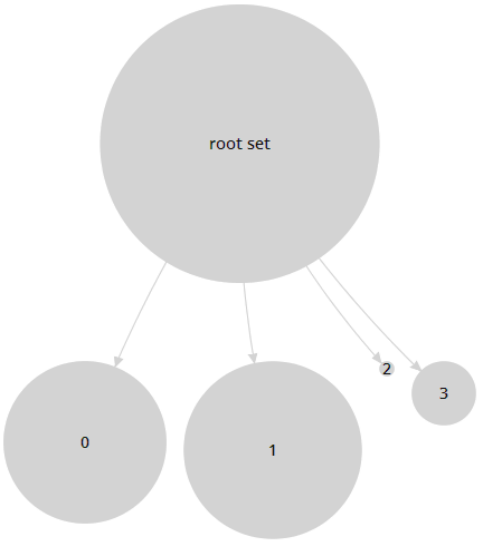


Figura 16. Gráfico de población de los clústeres

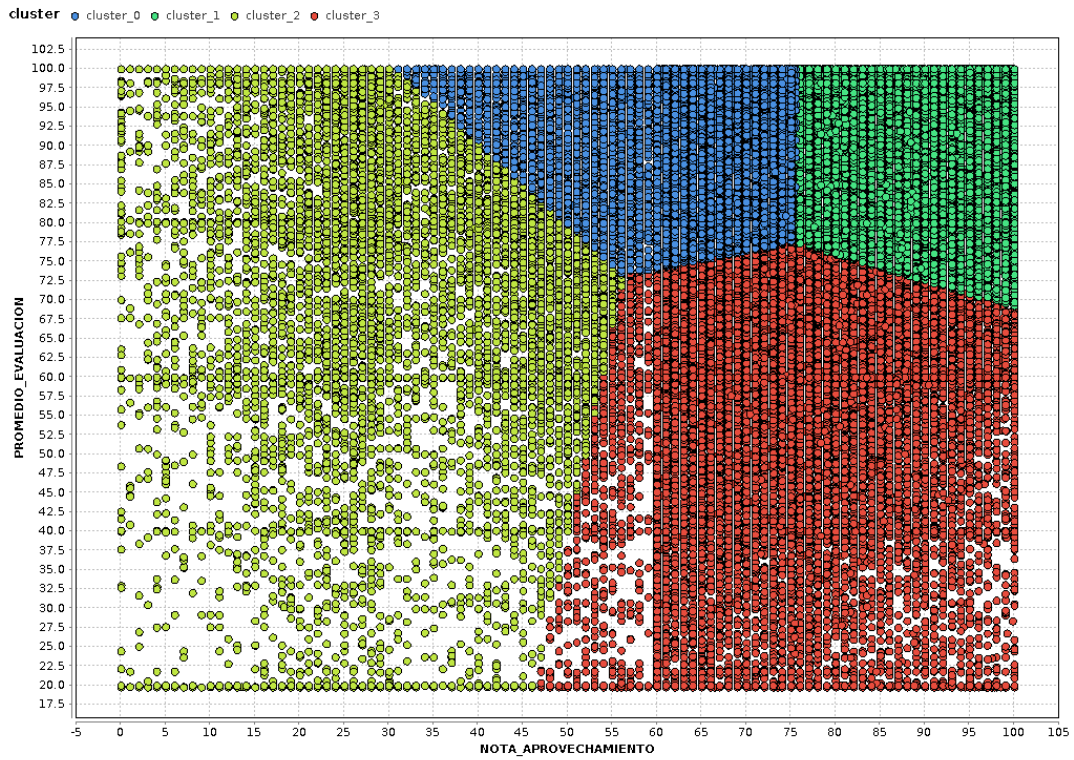


Figura 17. Separación de los grupos en 2D, eje x: aprovechamiento, eje y: evaluación docente, color: clústeres

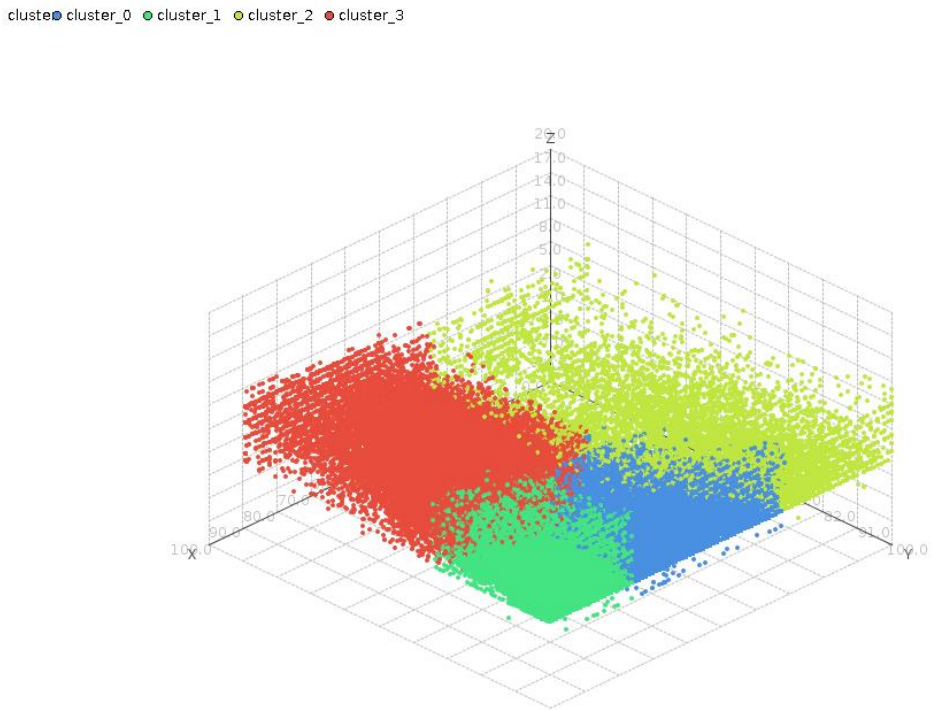


Figura 18. Separación de los grupos en 3D, eje x: aprovechamiento, eje y: evaluación docente, eje z: nivel / ciclo, color: clústeres

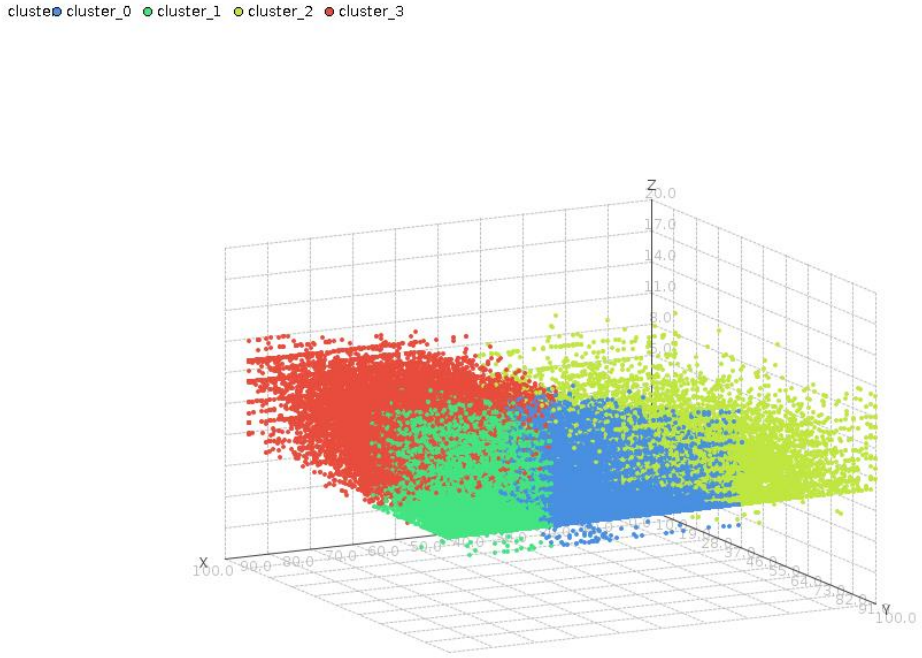


Figura 19. Separación de los grupos en 3D, eje x: aprovechamiento, eje y: evaluación docente, eje z: nivel / ciclo, color: clústeres

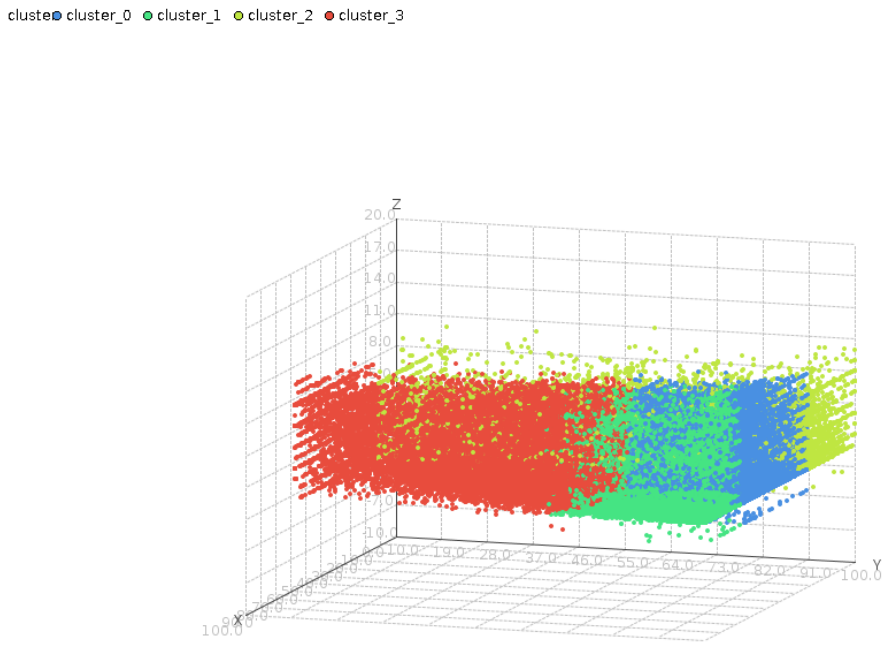


Figura 20. Separación de los grupos en 3D, eje x: aprovechamiento, eje y: evaluación docente, eje z: Nivel / Ciclo, color: clústeres

Proyecto:	Modelo predictivo de deserción estudiantil en estudiantes que ingresan a la universidad
Actividad:	**Evaluar los resultados**

Como se puede observar en la figura 16, los grupos o clústeres identificados poseen tamaños diferentes (tamaño hace referencia al número sujetos dentro de cada grupo), a continuación se presenta un listado por orden descendente con relación a su tamaño:

1. **cluster 1**: con 127.453 instancias, este clúster presenta un centroide ubicado en un aprovechamiento de 86 y una evaluación al docente de 84.
2. **cluster 0**: con 114.055 instancias, este clúster presenta un centroide ubicado en un aprovechamiento de 65 y una evaluación al docente de 94.
3. **cluster 3**: con 40.655 instancias, este clúster presenta un centroide ubicado en un aprovechamiento de 73 y una evaluación al docente de 58.
4. **cluster 2**: con 9.264 instancias, este clúster presenta un centroide ubicado en un aprovechamiento de 35 y una evaluación al docente de 58.

De la figura 17 se puede también observar que tanto el cluster_1 como el cluster_0, a pesar de poseer el mayor número de instancias son más compactos que los otros clústeres.

En base a los centroides identificados se procede a clasificar por niveles los puntajes tanto de aprovechamiento como de evaluación al docente:

- Aprovechamiento: alto (86), medio (73), regular (65) y muy bajo (35)
- Evaluación docente: alto (94), medio (84), bajo (58).

Aplicando los niveles anteriores sobre los clústeres formados, podemos identificar lo siguiente:

- Que el primer grupo (cluster_1), el de mayor tamaño, corresponde a los alumnos con aprovechamiento ***Alto*** que han calificado con una evaluación ***Media*** a sus docentes.
- Que el segundo grupo (cluster_0), corresponde a los alumnos con rendimiento ***Regular*** que han calificado con una evaluación ***Alta*** a sus docentes.
- Que el tercer grupo (cluster_3), corresponde a los alumnos con rendimiento ***Medio*** que han calificado con una evaluación ***Baja*** a sus docentes.
- Que el cuarto grupo (cluster_2) y el de menor tamaño, corresponde a los alumnos con rendimiento ***Muy Bajo*** que han calificado con una evaluación ***Baja*** a sus docentes.

Tabla 10. Ubicación de los clústeres con relación a los niveles obtenidos de los centroides.

	Evaluación Docente			
Aprovecha-miento		*Alto (94)*	*Medio (84)*	*Bajo (58)*
	Alto (86)		cluster_1 (*1*)	
	Medio (73)			cluster_3 (*3*)
	Regular (65)	cluster_0 (*2*)		
	Muy bajo (35)			cluster_2 (*4*)

ELABORACIÓN: Gustavo Cordero

A partir de los resultados obtenidos se puede interpretar lo siguiente:

- Que los alumnos con alto aprovechamiento (84) han evaluado con una calificación inferior a los docentes, con relación a la calificación dada por los alumnos de aprovechamiento medio (73), esto podría suponer que los últimos al poseer calificaciones cercanas al mínimo para aprobación (60), se sienten presionados a emitir una buena evaluación a sus docentes.
- Que existen dos grupos de alumnos que emiten bajas evaluaciones a sus docentes (58), los alumnos de aprovechamiento medio (73) y los de aprovechamiento muy bajo (35), los primeros podrían considerarse evaluaciones con criterio puesto que son alumnos promedio que si bien no tiene el aprovechamiento más alto, sus calificaciones les permiten aprobar sin problema sus materias, por lo que no tendrían presiones que afecten a su evaluación. No así el segundo grupo, ya que son alumnos cuyas calificaciones están por debajo del mínimo para aprobar las materias, es decir son alumnos que no aprobaron, por tanto la evaluación dada a sus docentes bajo este contexto serían muy subjetivas.

Inclusión de la variable Nivel al análisis:

De las figuras 18, 19 y 20 se puede observar que existe una tendencia: los alumnos son más estrictos en las evaluaciones según se encuentre en niveles superiores (evaluaciones con menor calificación), como se puede observar los grupos de cluster_1 y cluster_0 según aumentan los niveles estos pierden instancias, y a la inversa el grupo del cluster_3, con el aumento de niveles se visualiza también un aumento de instancias en este grupo.

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Evaluar el aporte del modelo**
El aporte del estudio realizado, es importante puesto que ha permitido determinar que si existe un agrupamiento natural de los datos correspondiente a la calificaciones de los estudiantes y la evaluación docente que ellos realizan y que además esta relación sufre ciertos cambios según el nivel en el que se encuentran los alumnos.	

Proceso para la gestión de proyectos de Minería de Datos - Universidad de Cuenca	
Proyecto:	Análisis mediante clustering de la relación entre el aprovechamiento estudiantil y la evaluación docente
Actividad:	**Revisar el proceso**
En la revisión al proceso se verifica la validez del estudio realizado puesto que se han seguido adecuadamente los lineamientos.	

6 CONCLUSIONES Y TRABAJOS FUTUROS

6.1 Conclusiones

Tras la ejecución del presente proyecto quedó claro que las técnicas de Minería de Datos, son herramientas de un extraordinario valor y poder para un oportuno análisis de datos. La Universidad de Cuenta por tanto debe hacer uso de estas técnicas de

manera continua y promover con ello una cultura de toma de decisiones basada en los datos (Data-Driven Decision Making). Empleando el proceso definido en el presente documento, se pudo evidenciar que el mismo es una guía acertada para no obviar actividades importantes dentro de un proyecto, a la vez que permiten mantener a la vista el objetivo de negocio a lo largo del proyecto. La evaluación realizada tras la primera iteración resultó muy útil para conocer aspectos claves de la problemática a estudiar y tener la posibilidad de priorizar los puntos críticos del proyecto. Sin embargo se pudo notar que las actividades de comprensión y selección de los datos son las actividades más demandantes del proceso por lo que contar con el personal que conozca el giro de negocio y los datos disponibles es fundamental. Una limitante importante relacionada a lo anterior fue que la implementación del sistema de Data Warehouse se encontró en un etapa muy temprana por lo que no se pudo utilizar como principal fuente de acceso a los datos, debiéndose tomar los datos de los propios sistemas transaccionales para cumplir con los objetivos planteados, esta situación debería superarse una vez que dicho sistema se encuentre en su etapa de madurez.

6.2 Trabajos futuros: oportunidades de información que potencien la aplicación de la Minería de Datos en estudios futuros

Como se mencionó en el Capítulo 2, la Minería de Datos es un punto de intersección de varias ciencias con el objetivo de generar información en base al análisis de grandes volúmenes de datos; su importancia dentro del sector educativo ha facilitado el aparecimiento de dos enfoques que buscan mejorar los sistemas educativos: EDM (Educational Data Mining) y LA (Learning Analytics); mientras EDM se enfoca en los aspectos técnicos del descubrimiento de conocimiento en información educacional, LA se enfoca en mejorar los aspectos de aprendizaje, por lo que se relaciona con campos más diversos como: educación, psicología, filosofía, sociología, lingüística, ciencias del aprendizaje, estadística y ciencias de la computación (Machine Learning, Inteligencia Artificial, Minería de Datos, etc.). Ambos enfoques presentan valiosos aportes para la aplicación de Minería de Datos en una institución educativa, en base a ello en la presente sección se va a proceder a identificar algunas de las principales aplicaciones de estos dos enfoques.

Los principales sistemas utilizados como fuente de datos en LA son los siguientes:

- Entorno Virtual de Aprendizaje o VLE por Virtual Learning Environment
- Sistema de información de estudiantes o SIS por Student Information Systems, comúnmente estos sistemas forman parte de los sistemas de gestión académica, poseen tanto información propia del estudiante como datos socioeconómicos, calificaciones previas, como también información de las materias tomadas, calificaciones obtenidas a la fecha, etc.

- Sistemas de Monitoreo de Asistencia, este tipo de sistemas permiten registrar y dar seguimiento a la asistencia de los estudiantes, incluso pueden permitir la identificación de estudiantes cuando ingresan en determinados sitios de la universidad como bibliotecas, laboratorios, etc., mediante uso de tarjetas de acceso.
- Sistemas de biblioteca: visitas del estudiante, libros solicitados, acceso a artículos digitales, etc.
- Aplicaciones Móviles: en este tipo de sistemas las aplicaciones son ilimitadas, se podrían construir desde herramientas que permitan a los estudiantes establecerse objetivos con relación a sus actividades escolares, y comprobar su progreso y resultados con otros estudiantes, así como evaluar continuamente a sus docentes o los contenidos recibidos en sus clases, o constituir herramientas para los docentes que permitan registrar el avance en sus contenidos, registrar observaciones sobre la planificación, evaluar a sus alumnos no sólo en rendimiento académico sino la participación dentro del aula, etc.

La integración entre estos sistemas posibilita la realización de interesantes análisis por ejemplo:

- En los sistemas académicos (SIS) se puede encontrar información sobre la situación financiera de los estudiantes, el número de horas que dedican a sus trabajos, sus rendimiento académico previo a que empiece a trabajar, etc. Esta información complementada por cuestionarios o encuestas desarrolladas por los estudiantes formarían una rica fuente de datos.
- Información de los sistemas de monitoreo de asistencia y tarjetas de acceso, podrían servir para analizar la responsabilidad o compromiso de los estudiantes.
- Los estudiantes no sólo utilizan los contenidos provistos por sus instituciones, ellos generan un gran cantidad de información en forma de: blogs, contenido audiovisual, programas de software, etc. Información que al ser digital es susceptible de ser considerada en LA, en este caso para medir la motivación del estudiante.
- Disponer de una plataforma que integre los datos producidos de las aplicaciones móviles mencionadas, generarían una información sumamente valiosa para analizar el proceso de enseñanza desde perspectivas tanto del docente como del alumno.

En la Figura 21, se presenta como ejemplo una arquitectura para Learning Analytics propuesta por Jisc, una iniciativa para dotar a las Universidades del Reino Unido de un servicio nacional de analítica de aprendizaje - LA. (Niall, 2016).

La arquitectura indicada propone la implementación de un Almacén de Registros de Aprendizaje (Learning records warehouse), el cual es alimentado por los sistemas de Biblioteca, VLE, SIS, e información auto-declarada de los alumnos. La parte medular de la arquitectura la conforma el Procesador de Analítica de Aprendizaje, donde el análisis predictivo es realizado, y da aviso al Sistema de Intervención y Alerta. La visualización de la analítica es posible gracias a la disposición de varios dashboards, y una aplicación móvil para estudiantes les permite visualizar sus datos y compararlos con otros alumnos. Por otro lado un servicio de consentimiento del estudiante ayuda a asegurar la privacidad permitiendo a los estudiantes dar sus permisos para la captura y uso de datos.

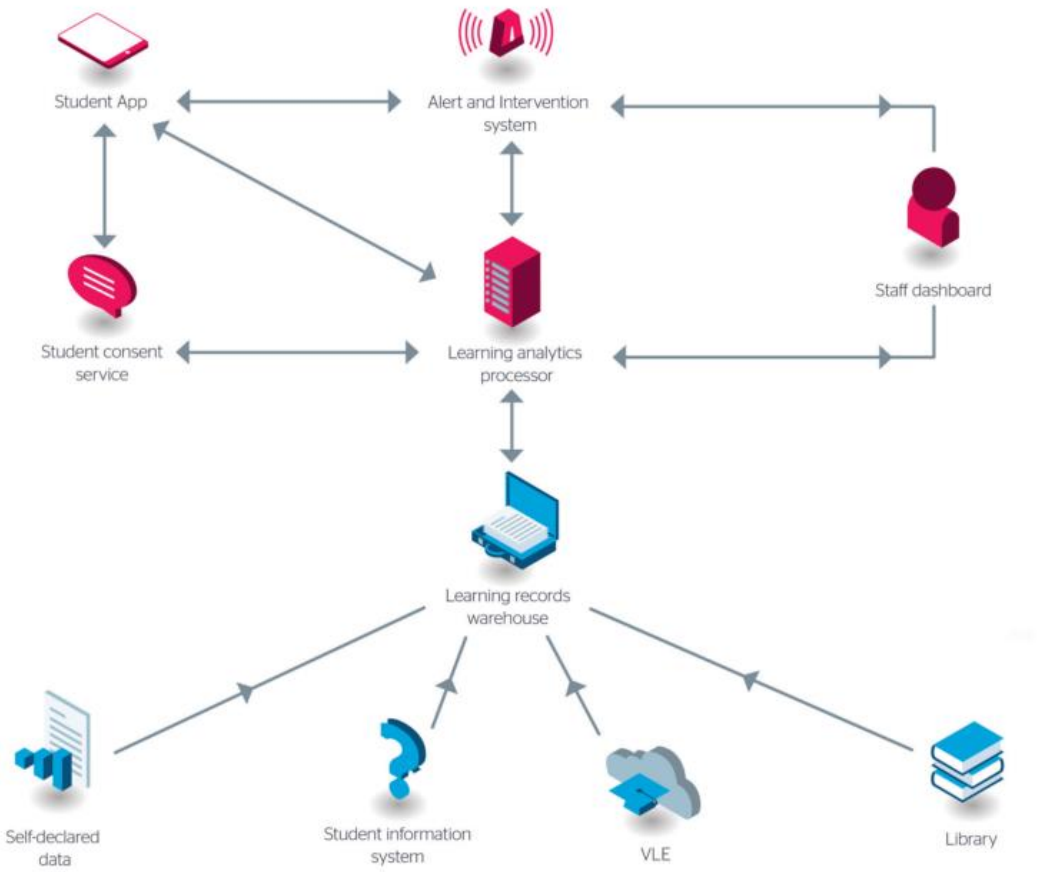


Figura 21. Arquitectura de Learning Analytics propuesta por JISC

Los sistemas fuente que se presenta en la arquitectura JISC son consistentes con los sistemas actuales de la universidad de Cuenca, por lo que la aplicación de esta arquitectura sería muy conveniente, y se aconseja iniciar con la implementación de la aplicación móvil mientras se integran y construye los demás componentes, esto permitiría recolectar datos históricos de una perspectiva que actualmente no es analizada por la universidad, el propio estudiante como centro de recolección de la información.

7 REFERENCIAS

Alnoukari, Mouhib, Zaidoun Alzoabi, and Saiid Hanna. 2008. “Applying Adaptive Software Development (ASD) Agile Modeling on Predictive Data Mining Applications: ASD-DM Methodology.” In 2008 International Symposium on Information Technology, 1–6. IEEE. doi:10.1109/ITSIM.2008.4631695.

Azevedo, Ana, Santos, Manuel, 2008. KDD, SEMMA and CRISP-DM: A Parallel Overview. *IADIS European Conference Data Mining*, no. January: 182–85.

Ballesteros, Alejandro, Daniel Sánchez-Guzmán, Ricardo García, 2013. Minería de Datos Educativa: Una Herramienta Para La Investigación de Patrones de Aprendizaje Sobre Un Contexto Educativo. *Latin-American Journal of Physics Education* 7 (4): 662–68.

Barreto, Silvia E, María V López, María G Ramírez Arballo, Eduardo A Porcel, and Liliana E

Chapman, Pete, Julián Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth, 2000. *Crisp-Dm 1.0.* CRISP-DM Consortium, 76.

Frawley, William, Piatetsky, Gregory, Matheus, Christopher, 1992. Knowledge Discovery in Databases: An Overview. *AI Magazine*. 13: 57-70.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 17(3).

Fischer, Erwin, 2012. *Modelo Para La Automatización Del Proceso de Determinación de Riesgo de Deserción En Alumnos Universitarios*, Tesis de Posgrado, Universidad de Chile, Santiago de Chile. Disponible en http://repositorio.uchile.cl/handle/2250/111188

Gorunescu, Florin, 2011). *Data Mining* (2ª ed.). San Francisco, CA, Morgan Kaufmann (Vol. 12). Springer-Verlag Berlin Heidelberg.

Gupta, Neha, 2013. Artificial Neural Network. *Network and Complex Systems* 3(1): 24–28.

Han, Jiawei, Kamber, Micheline, 2006. *Data Mining: Concepts and Techniques*. Annals of Physics (Vol. 54).

Moscoso, Oswaldo, Lujan, Sergio, 2016. *Minería de Datos Educacionales: una visión holística.* In 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–6.

Olson, David, Delen, Dursun (2008). Data Mining. In Advanced Data Mining Techniques (pp. 9–35). Springer.

Pacheco, Yoandry, Fernández, Yaima, 2015. Aplicación de técnicas de descubrimiento de conocimientos en el proceso de caracterización estudiantil. *Ciencias de la Información*, 46(3), 25-30

Patki, Priti Sudhir, Vishakha V. Kelkar, 2013. *Classification using different normalization techniques in Support Vector Machine.* International Conference on Communication Technology. (pp. 17-19).

Romero, Cristóbal, Ventura, Sebastián, Pechenizkiy, Mykola, Baker, Ryan, 2010. Handbook of Educational Data Mining. Chapman & Hall / CRC.

Reason, Robert, 2009. Student variables that predict retention: Recent research and new developments. *Naspa Journal* 46(3), 482-501.

Seidman, Alan, 2012. *College student retention: Formula for student success*. Maryland: Rowman and Littlefield Publishers, 295 pp.

Spositto, O., Etcheverry, M., Ryckeboer, H., Bossero, J. (2008). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil.

Baker, R.S.J.d., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1 (1), 3-17

Baker, R.S.J.d. (2010) Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, pp. 112-118. Oxford, UK: Elsevier.

Baker, R., Siemens, G. (2014) Educational data mining and learning analytics. In Sawyer, K. (Ed.) Cambridge Handbook of the Learning Sciences: 2nd Edition, pp. 253-274.

Cheng, Jiechao. 2017. “Data-Mining Research in Education,” no. March. http://arxiv.org/abs/1703.10117.

Beal, C.R., Qu, L., & Lee, H. 2006. Classifying learner engagement through integration of multiple data sources. Paper presented at the 21st National Conference on Artificial Intelligence (AAAI-2006), Boston, MA

Macfadyen, L. P., Dawson, S. 2010. Mining LMS data to develop an “early warning system” for educators: A proof of concept. Computers & Education. 588-599

Fancsali, S. (2012) Variable Construction and Causal Discovery for Cognitive Tutor Log Data: Initial Results. Proceedings of the 5th International Conference on Educational Data Mining, 238-239.

Romero, Cristobal, and Sebastián Ventura. 2010. “Educational Data Mining: A Review of the State of the Art.” IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 40 (6): 601–18. doi:10.1109/TSMCC.2010.2053532.

Bhagoriya, Nupur, and Priyanka Pande. 2017. “EDUCATIONAL DATA MINING IN THE FIELD OF HIGHER EDUCATION-A SURVEY” 6 (4): 697–99.

Li, Yan. 2014. “NEW ARTIFACTS FOR THE KNOWLEDGE DISCOVERY VIA DATA ANALYTICS (KDDA) PROCESS.” In Virginia Commonwealth University.

Fischer, Erwin Sergio. 2012. “Modelo Para La Automatización Del Proceso de Determinación de Riesgo de Deserción En Alumnos Universitarios,” 95. http://repositorio.uchile.cl/handle/2250/111188.

Business Process Model and Notation, (s.f) En Wikipedia. Recuperado el 17 de Agosto de 2017 de https://es.wikipedia.org/wiki/Business_Process_Model_and_Notation

Allweyer, Thomas. 2013. “Bpmn 2.0,” 0–24.

Sclater, Niall, Alice Peasgood, and Joel Mullan. 2016. “Learning Analytics in Higher Education,” no. April: 10. http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2016/analytics-in-higher-education.pdf.

Sclater, Niall, and Joel Mullan. 2017. “Jisc Briefing.” Effective Learning Analytics, no. January. http://repository.jisc.ac.uk/6560/1/learning-analytics_and_student_success.pdf.

ANEXO A. Traducción del modelo de referencia CRISP DM 1.0

A continuación se presenta una traducción parcial del Modelo de referencia CRISP DM 1.0 con el objetivo de permitir al lector profundizar sobre el modelo estándar de la industria.

Etapa 1: Comprensión del negocio:

- **Descripción:** Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, para luego convertir este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.
- **Tarea 1.1: Determinar los objetivos del negocio:**
  - **Descripción:** el primer objetivo del ***analista de datos*** es entender desde una perspectiva de negocio y dentro de un contexto, lo que el cliente quiere lograr. Se debe equilibrar correctamente los varios posibles objetivos y restricciones del cliente. El objetivo del analista de datos es descubrir factores importantes al inicio, y que podrían influir los resultados del proyecto. Se trata de hacer la pregunta correcta, la cual será contestada como resultado del proyecto.
  - **Salidas:**
    - **Salida 1.1.1: Antecedentes:** al inicio del proyecto es importante que se registre la información que se conoce sobre el negocio de la organización.
    - **Salida 1.1.2: Objetivos de negocio:** descripción del objetivo primario del cliente, desde una perspectiva de negocio. A parte de este objetivo primario, que otras preguntas de negocio relacionadas, le interesa al cliente analizar.
    - **Salida: 1.1.3 Criterios de éxito de negocio:** describa los criterios para un resultado satisfactorio y útil del proyecto, desde una perspectiva de negocio. Los criterios podrían ser de diversa naturaleza, tan específicos y capaces de ser medidos objetivamente como ser algo real o tal vez subjetivos como una recomendación, en caso de ser subjetivos debería indicarse quién dará el juicio subjetivo.

- **Tarea 1.2: Evaluar la situación:**
  - **Descripción:** si el objetivo de la tarea anterior era ponerse rápidamente al tanto de la situación, en esta tarea el analista de datos debe ampliarse sobre los detalles, por tanto involucra la investigación más detallada sobre todos los recursos, restricciones, presunciones y otros factores a considerarse en la determinación del objetivo de análisis de datos y plan de proyecto.
  - **Salidas:**
    - **Salida 1.2.1: Inventario de recursos:** listado de recursos disponibles para el proyecto, incluyendo el personal (Subject Matter Expert - SME, experto de datos, soporte técnicos, expertos en minería de datos), datos (extractos fijos, acceso en vivo, almacenes de datos - DWH, datos operacionales), recursos computacionales (plataformas de hardware), y software (herramientas de minería de datos, otros software relevantes).

- **Salida 1.2.2: Requerimientos, presunciones y restricciones:**
  - Lista de todos los requerimientos del proyecto, incluyendo: calendario de terminación, compresibilidad y calidad de los resultados, seguridad, aspectos legales (asegurar que el analista de datos tenga permitido el uso de los mismos).
  - Lista de presunciones hechas en el proyecto. Pueden ser sobre los datos y ser verificadas durante la minería de datos, pero también presunciones no verificables sobre el negocio relacionado con el proyecto. ***Es muy importante listar si estas presunciones comprometen la validez de los resultados.***
- **Salida 1.2.3: Riesgos y contingencias:**
  - Listar los riesgos o acontecimientos que podrían retrasar el proyecto o hacer que falle, y listar sus planes de contingencia (qué acción será tomada si estos riesgos o acontecimiento ocurren).
- **Salida 1.2.4: Terminología:**
  - Elaboración de un glosario de terminología relevante al proyecto:
    - Glosario de terminología relevante del negocio, apoya la comprensión del negocio disponible para el proyecto.
    - Glosario de terminología de minería de datos, ilustrada con ejemplos relevantes al problema de negocio en cuestión.
- **Salida 1.2.5: Costos y beneficios:** elaborar un análisis de costo - beneficio, que compare los gastos del proyecto con los beneficios potenciales al negocio si fuera exitoso.

- **Tarea 1.3: Determinar los objetivos de la minería de datos:**
  - **Descripción:** Un objetivo de minería de datos declara los objetivos de proyecto en términos técnicos. Por ejemplo, el objetivo de negocio podría ser “Aumentar ventas por catálogo a clientes existentes.” Un objetivo de minería de datos podrían ser “Predecir cuantas baratijas un cliente comprará, obteniendo datos de sus compras de tres años pasados, información demográfica (edad, sueldo, ciudad, etc.), y el precio del artículo.”
  - **Salidas:**
    - **Salida 1.3.1 Objetivos de la minería de datos:** describir las salidas esperadas del proyecto que permiten el logro de los objetivos de negocio.
    - **Salida 1.3.2 Criterios de éxito de la minería de datos:** definir los criterios para un resultado exitoso en términos técnicos. PE: cierto nivel de precisión en la predicción. Al igual que un criterio de éxito de negocio, éste puede ser necesario describirse en términos subjetivos, en cuyo caso el evaluador subjetivo debe ser identificado.

- **Tarea 1.4: Elabora el plan de proyecto:**

- **Descripción:** definir el plan adecuado para alcanzar los objetivos de la minería de datos y con ellos los objetivos de negocio. El plan debería especificar los pasos a realizar durante el resto del proyecto, incluyendo la selección inicial de herramientas y técnicas.
- **Salidas:**
  - **Salida 1.4.1 Plan de proyecto:** listar las etapas a ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias. Donde sea posible, haga explícito las iteraciones en gran escala en el proceso de minería de datos, por ejemplo: las repeticiones del modelado y las fases de evaluación. Como parte del plan de proyecto, es también importante analizar dependencias entre la planificación de tiempo y los riesgos. *El plan de proyecto es un documento dinámico en el sentido de que en el final de cada fase, son necesarios una revisión del progreso y logros y una actualización correspondiente del plan de proyecto es recomendado. Los puntos de revisión específicas para estas actualizaciones son parte del plan de proyecto.*

Etapa 2: Comprensión de los datos:

- **Descripción:** La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizarse primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.
- **Tarea 2.1 Recolectar datos iniciales**
  - **Descripción:** adquiera los datos listados como recursos del proyecto. La recolección inicial incluye la carga de dichos datos en alguna herramienta para compresión de datos. Este esfuerzo posiblemente conduce a los pasos iniciales de preparación de datos. Si los datos vienen de múltiples fuentes, la integración es una cuestión adicional, aquí o más tarde en las fases de preparación de datos.
  - Salidas:
    - **Salida 2.1.1 Informe de recolección de datos iniciales:** listado del conjunto de datos adquirido, sus ubicaciones, los métodos utilizados para su adquisición y cualquier problema encontrado. Registre los problemas y cualquier solución realizada. Esto servirá para futuras replicaciones, tanto de este proyecto como de proyectos similares.

- **Tarea 2.2 Describir los datos**
  - **Descripción:** examine las propiedades “gruesas” o “superficiales” de los datos e informe sobre los resultados.
  - **Salidas:**
    - **Salida 2.2.1 Informe de descripción de datos:** descripción de los datos adquiridos, incluyendo el formato de los mismo y la calidad de los datos (PE: número de registros y campos en cada tabla), identificación de los campos, y cualquier otra característica superficial

descubierta. Evaluar si los datos adquiridos satisfacen los requisitos relevantes.

- **Tarea 2.3 Explorar los datos**
  - **Descripción:** Esta tarea aborda las preguntas de minería de datos usando técnicas de consultas, visualización y reportes. Incluye la distribución de atributos claves relacionados entre pares o pequeños números de atributos, los resultados de simples agregaciones, las propiedades de las subpoblaciones significativas, y análisis estadísticos simples. Estos análisis directamente pueden dirigir los objetivos de minería de datos; ellos también pueden contribuir o refinar la descripción de datos e informes de calidad, y alimentar en la transformación y otros pasos de preparación de datos necesarios para análisis futuros.
  - **Salidas:**
    - **Salida 2.3.1 Informe de exploración de datos:** descripción de los resultados de esta tarea, incluyendo primeras conclusiones o hipótesis iniciales y su impacto sobre el resto del proyecto. Si es factible debe incluirse gráficos y plots para indicar las características de los datos que sugieran exámenes adicionales de subconjuntos de datos de interés.

- **Tarea 2.4 Verificar la calidad de los datos**
  - **Descripción:** examen de la calidad de los datos, preguntando cuestiones tales como:
    - ¿son los datos completos? (cubren todos los casos requeridos)
    - ¿Estos son correctos, o contienen errores?, si es así, ¿qué tan frecuentes son estos errores?
    - ¿Existen valores perdidos en los datos?, si es así, ¿Cómo están ellos representados?, ¿dónde esto ocurre?, ¿qué tan común estos son?
  - **Salidas:**
    - **Salida 2.4.1 Informe de calidad de datos:** consiste en un listado de los resultados de la verificación, si un problema de calidad existe, es adecuado listar las posibles soluciones. Las soluciones a los problemas de calidad de datos dependen profundamente de tanto del conocimiento de los datos como del negocio.

Etapa 3: Preparación de los datos:

- **Descripción:** La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos en las herramientas de modelado) de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescrito. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.
- **Salidas:** esta etapa posee dos salidas generales:

- **Conjunto de datos (dataset):** corresponde al conjunto o conjuntos de datos producidos a lo largo de esta fase que será usada para modelar o para el trabajo principal de análisis del proyecto.
- **Descripción del Conjunto de Datos (Dataset):** descripción de los conjuntos de datos que serán utilizados para el modelamiento y el trabajo principal de análisis del proyecto.

- **Tarea 3.1 Seleccionar los datos**
  - **Descripción:** Decidir qué datos serán usados para el análisis. Los criterios incluyen la importancia a los objetivos de la minería de datos, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Note que *la selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla.*
  - **Salidas:**
    - **Salida 3.1.1 Razonamiento para la inclusión o exclusión:** Listar los datos para ser incluidos/excluidos y los motivos para estas decisiones.
- **Tarea 3.2 Limpiar los datos**
  - **Descripción:** Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de los subconjuntos de datos limpios, la inserción de datos por defectos adecuados, o técnicas más ambiciosas tales como la estimación de datos faltantes mediante modelado.
  - **Salidas:**
    - **Salida 3.2.1 Informe de la limpieza de los datos:** describir qué decisiones y acciones fueron tomadas para dirigir los problemas de calidad de datos informados durante la tarea de ***Verificación de Calidad de Datos de los Datos*** de la fase de ***Comprensión de Datos***. Las transformaciones de los datos para una apropiada limpieza y el posible impacto en el análisis de resultados deberían ser considerados.
- **Tarea 3.3 Construir los datos**
  - **Descripción:** Esta tarea incluye la construcción de operaciones para preparación de datos tales como la producción de atributos derivados o el ingreso de nuevos registros, o la transformación de valores para atributos existentes.
  - **Salidas:**
    - **Salida 3.3.1 Atributos derivados:** los atributos derivados son los atributos nuevos que son construidos de uno o más atributos existentes en el mismo registro. Ejemplo: área = longitud * anchura.
    - **Salida 3.3.2 Registros generados:** describa la creación de registros completamente nuevos. Ejemplo: Crear registros para los clientes quienes no hicieron compras durante el año pasado. No había ninguna razón de tener tales registros en los datos brutos, pero para el objetivo del modelado esto podría tener sentido para representar explícitamente el hecho que ciertos clientes no hayan hecho compra nada.
- **Tarea 3.4 Integrar datos**

- **Descripción:** estos son los métodos por el cual la información es combinada de múltiples tablas o registros para crear nuevos registros o valores.
- **Salidas:**
  - **Salida 3.4.1 Datos combinados:** La combinación de tablas se refiere a la **unión simultánea de dos o más tablas** que tienen información diferente sobre el mismo objeto. Ejemplo: una cadena de venta al público tiene una tabla con la información sobre las características generales de cada tienda (Por ejemplo, el espacio, el tipo de comercio), otra tabla con datos resumidos de las ventas (por ejemplo, el beneficio, el cambio porcentual en ventas desde el año anterior), y el otro con información sobre los datos demográficos del área circundante. Cada una de estas tablas contienen un registro para cada tienda. Estas tablas pueden ser combinadas simultáneamente en una nueva tabla con un registro para cada tienda, combinando campos de las tablas fuentes. Los datos combinados **también cubren agregaciones**. La agregación se refiere a operaciones en la que nuevos valores son calculados de información resumida de múltiples registros y/o tablas. Por ejemplo, convirtiendo una tabla de compra de clientes donde hay un registro para cada compra en una tabla nueva donde hay un registro para cada cliente, con campos tales como el número de compras, el promedio de la cantidad de compra, el porcentaje de órdenes cobrados a tarjeta de crédito, el porcentaje de artículos bajo promoción, etc.

- **Tarea 3.5 Formatear Datos**
  - **Descripción:** las transformaciones de formato se refiere a modificaciones principalmente sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.
  - **Salidas:**
    - **Salida 3.5.1 Datos reformateados:** algunas herramientas tienen requerimientos sobre el orden de los atributos, tales como el primer campo que es un único identificador para cada registro o el último campo es el campo resultado que el modelo debe predecir. Podría ser importante cambiar el orden de los registros en el conjunto de datos. Quizás la herramienta de modelado requiere que los registros sean clasificados según el valor del atributo de resultado. Comúnmente, los registros del conjunto de datos son ordenados al principio de algún modo, pero el algoritmo que modela necesita que ellos estén en un orden moderadamente arbitrario. Por ejemplo, cuando se usa redes neuronales, esto es generalmente mejor para los registros para ser presentados en un orden aleatorio, aunque algunas herramientas manejen esto automáticamente sin la intervención explícita del usuario. Además, hay cambios puramente sintácticos hechos para satisfacer las exigencias de la herramienta de modelado específica. Ejemplos: al eliminar las comas dentro de los campos de texto en ficheros de datos delimitados por coma, corta todos los valores a un máximo de 32 caracteres.

Etapa 4: Modelado:

- **Descripción:** En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.
- **Tarea 4.1 Escoger la técnica de modelado**
  - **Descripción:** Como primer paso en modelado, seleccionar la técnica de modelado real que está por ser usado. Aunque usted haya podido seleccionar una herramienta durante la fase de Comprensión del negocio, esta tarea se refiere a la técnica de modelado específico, por ejemplo, un árbol decisión construido con C4.5, o la generación de red neuronales Back-Propagación. Si múltiples técnicas son aplicadas, se realizan esta tarea separadamente para cada técnica.
  - **Salidas:**
    - **Salida 4.1.1 Técnica de modelado:** documente la técnica de modelado real que está por ser usado.
    - **Salida 4.1.2 Presunciones de modelado:** muchas técnicas de modelado hacen presunciones específicas sobre los datos, por ejemplo: que todos los atributos tengan distribuciones uniformes, no encontrar valores no permitidos, el atributo de clase debe ser simbólico, etc. Registrar cualquiera de tales presunciones hechas.

- **Tarea 4.2 Generar la prueba de diseño:**
  - **Descripción:** antes de que nosotros en realidad construyamos un modelo, tenemos que generar un procedimiento o el mecanismo para probar la calidad y validez del modelo. Por ejemplo, en tareas de minería de datos supervisados como la clasificación, esto es común usar tasas de errores como medida de calidad para modelos de minería de datos. Por lo tanto, típicamente separamos el conjunto de datos en una serie y en un conjunto de prueba, construimos el modelo sobre el conjunto de series, y estimamos su calidad sobre el conjunto de prueba separado.
  - **Salidas:**
    - **Salida 4.2.1 Prueba de diseño:** describir el plan intencionado para el entrenamiento, la prueba, y la evaluación de los modelos. Un componente primario del plan determina como dividir un conjunto de datos disponible en datos de entrenamiento, datos de prueba, y conjunto de datos de validación.
    -
- **Tarea 4.3 Construir el modelo**
  - **Descripción:** Ejecutar la herramienta de modelado sobre el conjunto de datos preparados para crear uno o más modelos.
  - **Salidas:**
    - **Salida 4.3.1 Configuración de parámetros:** Con cualquier herramienta de modelado, hay a menudo un gran número de

parámetros que pueden ser ajustados. Listar los parámetros y sus valores escogidos, también con el razonamiento para elegir los parámetros de ajustes

- **Salida 4.3.2 Modelos:** Corresponde a los modelos reales producidos por la herramienta de modelado, no un informe.
- **Salida 4.3.3 Descripción de modelos:** Describir los modelos obtenidos. Informar sobre la interpretación de los modelos y documentar cualquier dificultad encontrada con sus significados.
-

- **Tarea 4.4 Evaluar el modelo**
  - **Descripción:** El ingeniero de minería de datos interpreta los modelos según su conocimiento de dominio, los criterios de éxitos de minería de datos, y el diseño de prueba deseado. El ingeniero de minería de datos juzga el éxito de la aplicación del modelado y descubre técnicas más técnicamente; él se pone en contacto con analistas de negocio y expertos en el dominio luego para hablar de los resultados de la minería de datos en el contexto de negocio. Por favor note que esta tarea sólo se considera modelos, mientras que la fase de evaluación también toma en cuenta todos los otros resultados que fueron producidos en el curso del proyecto.
    El ingeniero de minería de datos intenta clasificar los modelos. Él evalúa los modelos según los criterios de evaluación. Tanto como es posible, él también tiene en cuenta objetivos del negocio y criterios de éxito de negocio. En los grandes proyectos de minería de datos, el ingeniero de minería de datos aplica una sola técnica más de una vez, o genera resultados de minería de datos con varias técnicas diferentes. En esta tarea, él también compara todos los resultados según los criterios de evaluación
  - **Salidas:**
    - **Salida 4.4.1 Evaluación de modelos:** Resumir los resultados de esta tarea, listar las calidades de los modelos generados (por ejemplo, en términos de exactitud), y clasificar su calidad en relación con cada otro.
    - **Salida 4.4.2 Parámetros de ajuste revisados:** Según la evaluación del modelo, revise los parámetros de ajuste y afínelos para la siguiente ejecución de la tarea de **Construir el Modelo**. Repetir la construcción y evaluación del modelo hasta que crea que usted ha encontrado el/los mejor/es modelo/s. Documentar todo como las revisiones y las evaluaciones.

Etapa 5: Evaluación:

- **Descripción:** en esta etapa ya se ha construido un modelo (o modelos) que parece tener alta calidad desde una perspectiva de análisis de datos. Se requiere evaluar a fondo el modelo y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, se debería obtener una decisión en el uso de los resultados de minería de datos.
- **Tarea 5.1 Evaluar los resultados**

- **Descripción:** Los pasos previos de la evaluación tratan con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado en que el modelo cumple los objetivos de negocio y busca determinar si hay alguna razón de negocio por la cual este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.
- Además, la evaluación también verifica otros resultados generados por la minería de datos. Los resultados de la minería de datos implican modelos que necesariamente son relacionados con los objetivos originales de negocio y todas los otros descubrimientos que no son relacionados necesariamente con los objetivos originales de negocio, pero también podría revelar desafíos adicionales, información, o insinuaciones para futuras direcciones.
- **Salidas:**
  - **Salida 5.1.1 Evaluación de los resultados de la minería de datos en lo que concierne a criterios de éxito del negocio:** Resumir los resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final en cuanto si el proyecto ya encuentra los objetivos iniciales de negocio.
  - **Salida 5.1.2 Modelos aprobados:** después de la evaluación de modelos en lo que concierne a criterios de éxito de negocio, los modelos generados que encuentran los criterios seleccionados son los modelos aprobados.

- **Tarea 5.2 Revisar el proceso**
  - **Descripción:** En este punto, los modelos resultantes parecen ser satisfactorios y satisfacer las necesidades del negocio. Ahora es apropiado hacer una revisión más completa de la minería de datos para determinar si hay cualquier factor importante o tarea que de algún modo ha sido pasada por alto. Esta revisión también cubre cuestiones de calidad, por ejemplo: ¿Construimos correctamente el modelo? ¿Usamos sólo los atributos que nos permitieron usar y que están disponibles para análisis futuros?
  - **Salidas:**
    - **Salida 5.2.1 Revisión del proceso:** resumir la revisión de proceso y destacar las actividades que han sido omitidas y/o aquellas que deberían ser repetidas.
- **Tarea 5.3 Determinar los próximos pasos**
  - **Descripción:** Dependiendo de los resultados de la evaluación y la revisión del proceso, el equipo de proyecto decide cómo proceder. El equipo decide si hay que terminar este proyecto y pasar a la implementación, inicia nuevas iteraciones o preparar nuevos proyectos de minería de datos. Esta tarea incluye el análisis de recursos restantes y del presupuesto, que puede influir en las decisiones.
  - **Salidas:**
    - **Salida 5.3.1 Lista de posibles acciones:** Listar las acciones potenciales futuras, con los motivos a favor y en contra de cada opción.
    - **Salida 5.3.2 Decisión:** Describa la decisión sobre cómo proceder, junto con la justificación.

Etapa 6: Despliegue - Desarrollo:

- **Descripción:** En ocasiones la creación per se dé un modelo no es el final del proyecto, sino que el conocimiento obtenido a través de este modelo, debería ser organizado y presentado de manera que el cliente pueda utilizarlo. Esto en ocasiones implica la aplicación de modelos "vivos" dentro de un ***proceso de toma de decisiones*** de la organización. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la organización. Muchas veces es el cliente, y no el analista de datos, quien lleva el paso de desarrollo.
- **Tarea 6.1 Desarrollar el plan:**
  - **Descripción:** Esta tarea toma los resultados de la evaluación y determina una estrategia para el despliegue. Si un procedimiento general ha sido identificado para crear el, o los modelos relevantes, este procedimiento es documentado aquí para el despliegue posterior.
  - **Salidas:**
    - **Salida 6.1.1 Plan de despliegue:** Resumir la estrategia de despliegue, incluyendo los pasos necesarios y cómo realizarlos.
- **Tarea 6.2 Planear la supervisión y mantenimiento:**
  - **Descripción:** La supervisión y el mantenimiento son cuestiones importantes si los resultados de la minería de datos son parte del negocio cotidiano y de su ambiente. La preparación cuidadosa de una estrategia de mantenimiento, ayuda a evitar largos e innecesarios periodos de uso incorrecto de los resultados de minería de datos. Para supervisar el despliegue de los resultados de la minería de datos, el proyecto necesita un plan detallado para el proceso de supervisión. Este plan tiene en cuenta el tipo específico de despliegue.
  - **Salidas:**
    - **Salida 6.2.1 Plan de monitoreo y mantenimiento:** Resumir la estrategia de supervisión y mantenimiento incluyendo los pasos necesarios y cómo realizarlos.
- **Tarea 6.3 Producir el informe final**
  - **Descripción:** al final del proyecto, el equipo del proyecto redacta un informe final. Según el tipo de despliegue, este informe puede ser sólo un resumen del proyecto y sus experiencias (si estas aún no han sido documentadas como una actividad en curso) o esto puede ser una presentación final y comprensiva de los resultados de minería de datos.
  - **Salidas:**
    - **Salida 6.3.1 Reporte Final:** Esto es el informe escrito final del resultado de la minería de datos. Este incluye todos los entregables anteriores, resumiendo y organizando los resultados.
    - **Salida 6.3.2 Presentación Final:** A menudo habrá también una reunión al cierre del proyecto, en el que los resultados son presentados verbalmente al cliente.
- **Tarea 6.4 Revisar el proyecto**

- **Descripción:** Evaluar lo que fue correcto y equivocado, lo que fue bien hecho y que necesita ser mejorado.
- **Salidas:**
  - **Salida 6.4.1 Documentación de la experiencia:** Resumir las experiencias importantes ganadas durante el proyecto. Por ejemplo, trampas, enfoques engañosos, o sugerencias para seleccionar las técnicas de minería de datos más adecuadas en situaciones similares podrían formar parte de esta documentación. En proyectos ideales, la documentación de la experiencia también cubre cualquier informe que ha sido escrito por miembros individuales del proyecto durante las fases del proyecto y sus tareas.